Studienstiftung des deutschen Volkes

Topics in Least-Squares and Discontinuous Petrov-Galerkin Finite Element Analysis

DISSERTATION

zur Erlangung des akademischen Grades
Dr. rer. nat.
im Fach Mathematik

von

Johannes Storn

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

Präsidentin der Humboldt-Universität zu Berlin:
Prof. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

1. Gutachter: Prof. Dr. Carsten Carstensen
2. Gutachter: Prof. Dr. Gerhard Starke
3. Gutachter: PD Dr. Dietmar Gallistl

Tag der Verteidigung: 5. Juli 2019

Zusammenfassung

Aufgrund der fundamentalen Bedeutung partieller Differentialgleichungen zur Beschreibung von Phänomenen in angewandten Wissenschaften ist deren Analyse ein Kerngebiet der Mathematik. Durch Computer lassen sich die Lösungen für eine Vielzahl dieser Gleichungen näherungsweise bestimmen. Die dabei verwendeten numerischen Verfahren sollen auf möglichst exakte Approximationen führen und deren Genauigkeit verifizieren. Die Least-Squares Finite-Elemente-Methode (LSFEM) und die unstetige Petrov-Galerkin (DPG) Methode sind solche Verfahren. Sie werden in dieser Dissertation untersucht.

Der erste Teil der Arbeit untersucht die Genauigkeit der mittels LSFEM berechneten Näherungen. Dazu werden Eigenschaften der zugrundeliegenden partiellen Differentialgleichungen mit den Eigenschaften der LSFEM kombiniert. Dies zeigt, dass die Abweichung der berechneten Näherung von der exakten Lösung einem berechenbaren Residuum asymptotisch entspricht. Ferner wird ein neues Verfahren zu Berechnung einer garantierten oberen Fehlerschranke eingeführt. Während der etablierte Fehlerschätzer den Fehler signifikant überschätzt, zeigen numerische Experimente eine äußerst geringe Überschätzung des Fehlers mittels der neuen Fehlerschranke.

Die Analyse der Fehlerschranken für das Stokes-Problem offenbart ein Beziehung der LSFEM und der Ladyschenskaja-Babuška-Brezzi (LBB) Konstanten. Diese Konstante ist entscheidend für die Existenz und Stabilität von Lösungen in der Strömungslehre. Der zweite Teil der Arbeit nutzt diese Beziehung und entwickelt ein auf der LSFEM basierendes Verfahren zur numerischen Berechnung der LBB Konstanten.

Der dritte Teil der Arbeit untersucht die DPG Methode. Dabei werden existierende Anwendungen der DPG Methode in einem abstrakten Rahmen zusammengefasst und analysiert. Diese Analyse zeigt, dass sich die DPG Methode als eine leicht gestörte LSFEM interpretieren lässt. Diese Interpretation erlaubt die Anwendung der Resultate aus dem ersten Teil der Arbeit und ermöglicht dadurch eine genauere Untersuchung existierender und die Entwicklung neuer DPG Methoden.

Abstract

The analysis of partial differential equations is a core area in mathematics due to the fundamental role of partial differential equations in the description of phenomena in applied sciences. Computers can approximate the solutions to these equations for many problems. They use numerical schemes which should provide good approximations and verify the accuracy. The least-squares finite element method (LSFEM) and the discontinuous Petrov-Galerkin (DPG) method satisfy these requirements. This thesis investigates these two schemes.

The first part of this thesis explores the accuracy of solutions to the LSFEM. It combines properties of the underlying partial differential equation with properties of the LSFEM and so proves the asymptotic equality of the error and a computable residual. Moreover, this thesis introduces an novel scheme for the computation of guaranteed upper error bounds. While the established error estimator leads to a significant overestimation of the error, numerical experiments indicate a tiny overestimation with the novel bound.

The investigation of error bounds for the Stokes problem visualizes a relation of the LSFEM and the Ladyzhenskaya-Babuška-Brezzi (LBB) constant. This constant is a key in the existence and stability of solution to problems in fluid dynamics. The second part of this thesis utilizes this relation to design a competitive numerical scheme for the computation of the LBB constant.

The third part of this thesis investigates the DPG method. It analyses an abstract framework which compiles existing applications of the DPG method. The analysis relates the DPG method with a slightly perturbed LSFEM. Hence, the results from the first part of this thesis extend to the DPG method. This enables precise investigations of existing and the design of novel DPG schemes.

Contents

1 Introduction

Minimal residual methods, like the least-squares finite element method (LSFEM) and the discontinuous Petrov-Galerkin (DPG) method, solve challenging partial differential equations by the minimization of an (artificial) energy. The minimization problem is in a Rayleigh-Ritz-like environment and so shares the advantageous mathematical and algorithmic properties of the well-understood Rayleigh-Ritz setting. The resulting numerical schemes are competitive for numerous problems of practical interest like fluid flows, elasticity, and convection-diffusion. This thesis exploits the underlying structure of the Rayleigh-Ritz-like environment and so contributes to the following topics.

1.1 Topics of this thesis

Error control for LSFEM. Many problems in applied sciences require fully reliable modelling, that is, they require an efficient approximation of the solution to a partial differential equation and a reliable verification of the accuracy. LSFEMs provide a built-in error control which allows for both aspects [BG09]: it drives adaptive mesh refinement algorithms (which capture singularities of the solution, see for example the triangulation of the Fichera Corner domain in Figure 1.1) and it leads to guaranteed upper error bounds (GUBs). This thesis proves asymptotic exactness of the built-in error control and improves the GUBs. The convergence history plot in Figure 1.1 visualizes both aspects, that is, it shows that the built-in error control (—♦—) approaches the error (—●—) in the log-log plot (and so the ratio of the error control and the error tends to one) and the novel guaranteed error bound (—○—) improves the guaranteed error bound (—■—) by several orders of magnitude as the mesh is refined.

Analysis of DPG. The discontinuous Petrov-Galerkin (DPG) method is a novel minimal residual method which has attracted much attention. The method is easy to implement, highly parallelizable, and allows for irregular triangulations with curved elements. The verification of three hypotheses provides a built-in error control and instant stability [CDG14, GQ14]. Carstensen, Demkowicz, and Gopalakrishnan verify the hypotheses for a large class of problems in [CDG16]. This thesis extends their results. This extension includes the definition and analysis of non-standard traces like traces of operators from parabolic and hyperbolic problems. The analysis circumvents the splitting in [CDG16, Thm. 3.1] and so improves the insight into the underlying structures. This allows for sharp estimates of the inf-sup constants β which enter GUBs. The (sharp) estimate of β in this thesis improves the state-of-the-art estimates β_{CDG} from [CDG16, Thm. 3.1] and β_{CP} from [CP18, Thm. 3.3]. For example, the improvement for the Poisson model problem on the unit square domain reads (see Remark 5.2.4 for more details)

$$\beta_{\mathrm{CDG}} = 0.441 < \beta_{\mathrm{CP}} = 0.607 < \beta = 0.833.$$

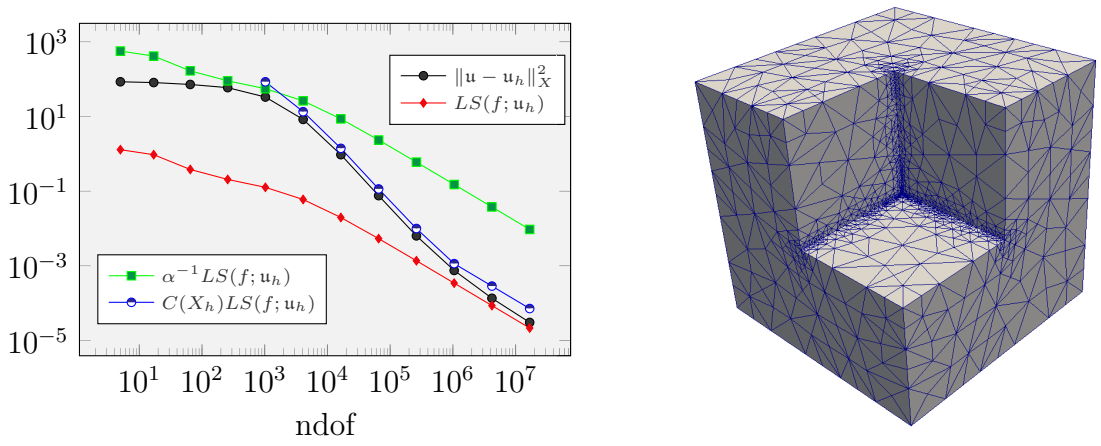


Figure 1.1: Convergence history plot of the error $\|u - u_h\|_X^2$, residual $LS(f; u_h)$, natural GUB $\alpha^{-1} LS(f; u_h)$, and improved GUB $C(X_h) LS(f; u_h)$ for the Helmholtz equation from [CS18, Fig. 1] and the adaptively refined mesh from Experiment 4 on page 40 for the Maxwell equations

The improved knowledge of the inf-sup and continuity constant allows to introduce optimally weighted test norms. These weighted test norms lead to optimal stability constants in the a priori error analysis. In addition, the analysis of the DPG method in this thesis links the DPG method with a (perturbed) LSFEM. This enables the application of results for LSFEMs to DPG methods. This thesis utilizes this observation to extend the asymptotic exactness results for LSFEMs to the primal DPG method for the Helmholtz equation and to design a locking-free primal DPG method for linear elasticity.

Computation of the LBB constant. The Ladyzhenskaya-Babuška-Brezzi (LBB) constant C_{LBB} is a key in the mathematical analysis of fluid dynamics and related problems [CCDL15]. Its value enters reliability constants, influences the convergence rate of iterative algorithms and allows the computation of the Babuška-Aziz constant, the Friedrichs constant for conjugate harmonic functions, and the constant in Korn's second inequality. Since the LBB constant is the smallest non-zero element in a non-compact eigenvalue problem, its approximation is challenging. This thesis introduces a competitive numerical scheme with monotonically convergent approximations $C_{\mathrm{LBB},h} \searrow C_{\mathrm{LBB}}$ as the maximal mesh-size of the underlying triangulation tends to zero.

1.2 State of the art

Error control for LSFEM. The number of fully reliable a posteriori error estimators is huge and includes explicit residual-based [BHHW00, CM14, MW01], averaging [BC02, CB02, BC04, BV00, Car99, Car02, Car04, CA03, CF01a, CF01c, CF01b, CV99, Rod94, ZZ87], equilibrated [Bra07, BMS, BS08, CM13, EV15], localized [CF99, CF00, PSD09], and functional [HHNL88, Rep97, Rep98, Rep99a, Rep99b, Rep00, RSS03, RX96] error estimators. The latter estimator is closely related to minimal residual methods in the sense that the minimization of the functionals results in a minimal residual method and vice versa any residual which is equivalent to the error with known equivalence constants

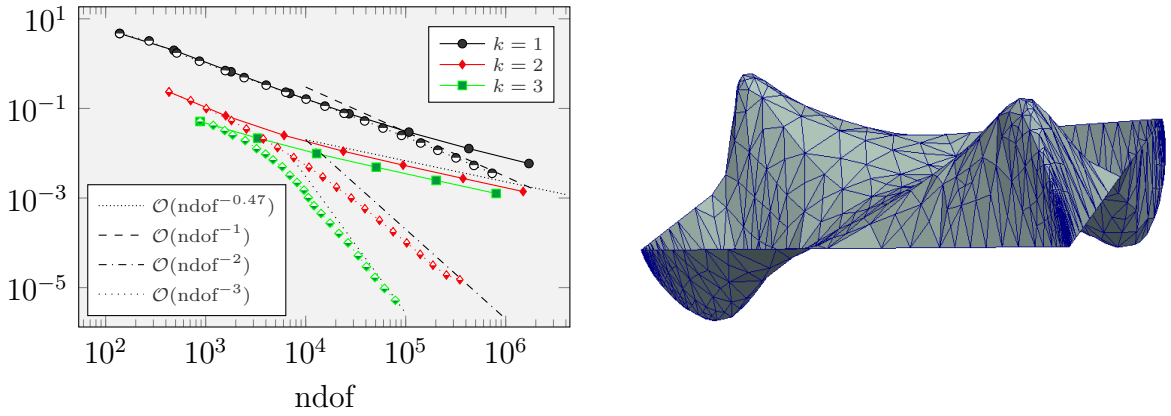


Figure 1.2: Convergence history plot of $(C^2_{\mathrm{LLB},h} - C^2_{\mathrm{LBB}})/C^2_{\mathrm{LBB}}$ with uniform (solid line) and adaptive (dotted line) mesh refinements and the y-component of the eigenfunction $\phi_h = (\phi_{h,x}, \phi_{h,y}) \in S^3_0(\mathcal{T}; \mathbb{R}^2)$ from Experiment 2 on page 56

leads to a fully reliable functional error estimator. However, the arguments in the analysis of LSFEMs for the Poisson model problem [CLMM94, JP93, PC94, PCL94], linear elasticity [CS03, CS04, SSS11], the Maxwell equations [BKP05b], and the Stokes problem [BG94, CLW04, CY00, SY97] lead to significant overestimations of the exact equivalence constants. This results in inefficient guaranteed error bounds. Nevertheless, the local contributions of the least-squares residuals drive adaptive mesh refinement algorithms. Numerical experiments [BMM97, BCCO09, CKS05, FMM98, Sta07] suggest that these adaptive schemes lead to optimal convergence rates. The only existing convergence proof of these algorithms [CPB17] requires a sufficiently large bulk parameter and so does not lead to an optimal convergence rate in general. Instead, optimal convergence proofs (see [CFPP14] for a detailed discussion on optimally convergent algorithms) for adaptive least-squares finite element methods [BC17a, BC17b, BCS18, CP15, CR17] utilize alternative residual-based error estimators and the separate marking strategy from [BM08, CR11].

Analysis of DPG. Demkowicz and Gopalakrishnan suggested some costly spatial decomposition of test functions in a framework of a minimal residual method with more test functions than seemingly necessary in their idealized DPG paradigm [DG10, DG11, DGN12, ZMD+11]. A modification in [GQ14] leads to the practical DPG method, which applies to a broad class of problems. An abstract functional analytical framework is based on the existence of a Fortin operator or an equivalent discrete inf-sup condition [CGHW14, CH16] and proves a priori [GQ14] as well as a posteriori error estimates [CDG14]. Carstensen's, Demkowicz', and Gopalakrishnan's publication [CDG16] validates the functional analytical framework for a large class of problems, including the Stokes equation [CP18], the transport equation [BDS18], linear elasticity [CH16], the Helmholtz equation [PD17], the Maxwell equations [CDG16], and viscoelastic fluid flows [KKR+17]. Recent research directions include time-stepping methods [FHSG17], space-time discretizations [DGNS17, Ern18, EW19, GS17], weighted test spaces [GMO14], non-linear DPG methods [CBHW18, FHS18], and optimally convergent adaptive schemes [CH18, Hel18].

Computation of the LBB constant. The existence of a continuous right inverse of the

divergence operator is important in the analysis of fluid dynamics and related problems. Its proof dates back to Ladyzhenskaya [Lad63] and Babuška and Aziz [BA72]. Bogovskiĭ extended the result to arbitrary Lipschitz domains [Bog79] and Durán, Muschietti, and coauthors generalized Bogovskiĭ's approach for John domains [ADM06, Dur12, DMRT10]. Brezzi's fundamental paper on mixed variational formulations [Bre74] links the continuity constant of the right inverse, called Babuška-Aziz constant, with a inf-sup constant C_{LBB}, called Ladyzhenskaya-Babuška-Brezzi (LBB) constant. It is well known that the LBB constant is the smallest non-zero element in the spectrum of the Cosserat operator [Vel96, Vel98]. Horgan and Payne introduced further relations, namely the relation of the Babuška-Aziz constant with the constant in Korn's second inequaltiy and the constant in the Friedrichs inequality for conjugate harmonic functions in [HP83]. Costabel and Dauge generalized these relations for arbitrary domains [CD15]. These relations allow for a counterexample of Ladyzhenskaya's result for domains with an external cusp based on Friedrichs' work [Fri37]. The classes of domains with known LBB constant include balls, ellipsoids, annular domains, spherical shells, and some domains defined by simple conformal images of a disk (see [CCDL15] and the references therein), but do not include simple domains like the square. Costabel et al. introduced a numerical scheme for the approximation the LBB constant in [CCDL15] to investigate the value numerically. They compute a discrete inf-sup constant β_h and prove that any accumulation point of β_h must be less or equal to the inf-sup constant $\beta = C_{\mathrm{LBB}}$. A sufficient condition which implies the convergence of β_h towards β requires discretizations of the pressure and velocity on different meshes with vanishing ratio of their mesh-sizes. Recently, Gallistl replaces the H^{-1} norm by a discrete H^{-1} norm which behaves monotonically under mesh refinements and so leads to a much better convergent numerical scheme [Gal19].

1.3 Results

Error control for LSFEM. This thesis introduces an abstract setting for LSFEMs. This setting allows for a spectral analysis, which leads to computable equivalence constants of the error and the computable residual. Moreover, the setting implies

(a) an asymptotic exactness property, which states that the ratio of the computable residual and error tends to one as the maximal mesh-size tends to zero, and
(b) an asymptotic best approximation property, which states that the ratio of the error and the best approximation error tends to one as the maximal mesh-size tends to zero.

Since the natural guaranteed upper bound (GUB) equals the computable residual times an (often large) equivalence constant, the asymptotic exactness property implies inefficiency of the natural GUB for fine meshes. This thesis remedies this downside by

(c) an improved computable reliability constant.

This thesis emphasizes the generality of the abstract setting (and so (a)–(c)) by

(d) the verification of the abstract setting for the Poisson model problem, the Helmholtz equation, linear elasticity, the Maxwell equations, and the Stokes equation.

Numerical experiments visualize the asymptotic results (a)–(b) for the model problems and underline the efficiency of the improved GUB (c). Unfortunately, the abstract setting

for the Stokes equation requires non-standard discretizations. Indeed, numerical experiments suggest that the asymptotic results (a)–(b) do not apply for the Stokes problem with standard discretizations.

Analysis of DPG. This thesis develops an abstract framework which compiles primal, mixed, and ultra-weak DPG methods. The framework leads to well-defined traces on the skeleton and verifies well-posedness of the variational formulation. This enable the possibility to

(e) design and analysis DPG methods for a wide class of problems, including parabolic and hyperbolic problems.

The abstract framework relates the DPG method with the LSFEM. This relation allows to apply results from the LSFEM and so leads to

(f) the asymptotic properties (a)–(b) for primal DPG methods,
(g) DPG methods with improved stability due to weighted test norms,
(h) a locking-free primal DPG method for linear elasticity.

Numerical experiments underline the asymptotic properties (f) for the Helmholtz equation and investigate the locking free-primal DPG method (h).

Computation of the LBB constant. The analysis of the LSFEM for the Stokes equations in (d) relates the equivalence constants with the LBB constant. This relation and the Rayleigh-Ritz-like environment of the LSFEM result in

(i) a monotonically convergent numerical scheme for the computation of the LBB constant with standard finite element spaces and symmetric positive definite matrices.

Numerical experiments with adaptive mesh refinements indicate optimal convergence rates and show that the method is competitive.

1.4 Methodology

This thesis investigates the LSFEM and DPG method for various model problems. The large number of model problems indicates the generality and benefits of this investigation, but interferes with a brief presentation. This section emphasizes the main ideas and so guides through this thesis.

Combination of spectral decompositions and the Galerkin orthogonality. Spectral decompositions are a well-known tool for the computation of inf-sup constants, see for example [DV98, Thm. 1] and [Bar15, p. 101]. Section 3.2.2–3.2.3 utilizes this idea to compute the ellipticity constants in LSFEMs for the Poisson model problem, Helmholtz equation, linear elasticity, Maxwell equations, and Stokes problem. These constants enter a priori and a posteriori error estimates. It turns out that only a few eigenfunctions cause large reliability and stability constants. Often, the discrete space approximates these eigenfunctions very well. Section 3.1 combines the good approximation of the eigenfunctions by discrete eigenfunctions and the Galerkin orthogonality of the LSFEM to improve the reliability and stability constants. This ansatz leads to asymptotic exactness results

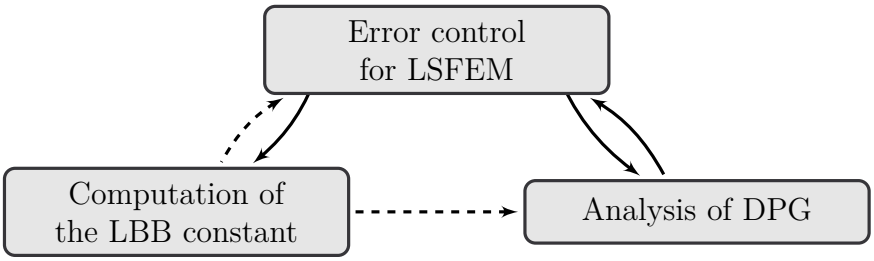


Figure 1.3: Relation of the three topics in this thesis (the dashed lines indicate that the LBB constant enters reliability and stability constants)

(Section 3.1.1) and an improved built-in error control (Section 3.1.2).

Approximate the LBB constant in a Rayleigh-Ritz-like environment. The LBB constant affects many stability constants in fluid dynamics. Thus, it is not surprising that the computations in Section 3.2.3 show a relation of the ellipticity constants in the LSFEM for the Stokes problem and the LBB constant. In contrast to the computation of non-compact eigenvalues, the computation of ellipticity constants in a Rayleigh-Ritz-like environment allows for the application of well-known results for the approximation of eigenvalues in a Rayleigh-Ritz setting. Chapter 4 exploits this observation to investigate the approximation of the LBB constant via the computation of the ellipticity constants. The result is a competitive numerical scheme which yields monotonically convergent approximations.

Extension of traces with useful properties. The analysis of DPG methods in Section 5.1.5 takes advantage of the fact that traces allow for very flexible extensions onto the interior of the domain. Carstensen, Demkowicz, and Gopalakrishnan exploit this possibility in [CDG16] by the definition of a trace extension as the solution to a partial differential equation with Dirichlet boundary conditions. In contrast to [CDG16], the characterization of the trace extension in Section 5.1.5 utilizes solely the Riesz representation theorem. This allows to introduce well-defined traces and so motivates a very general abstract setting for DPG methods in Section 5.1.4.

Exploit the relation of DPG and LSFEM. The trace extension in Section 5.1.5 relates the DPG method and the LSFEM. This relation allows to reuse the results from the LSFEM for the DPG method. This leads to computable and sharp inf-sup constants, optimal weighted test spaces for ultra-weak DPG methods (Section 5.1.6), the asymptotic exactness results (a)–(b) for the Helmholtz equation (Section 5.2.1), and a locking-free primal DPG method for linear elasticity (Section 5.2.2).

Acknowledgements

The author thanks Professor Dr. Carsten Carstensen for the supervision of this thesis, PD Dr. Dietmar Gallistl for his useful hints, and the workgroup Numerical Analysis at the Humboldt Universität zu Berlin for interesting and helpful discussions. Moreover, the author thankfully acknowledges the support of the Studienstiftung des deutschen Volkes.

2 Model problems

This thesis introduces concepts for the analysis of least-squares and discontinuous Petrov-Galerkin methods. These concepts apply to a broad range of important linear problems, among others the Poisson model problem, the Helmholtz equation, linear elasticity, the Maxwell equations, and the Stokes problem. This preliminary chapter introduces these five model problems and thereby recalls important textbook results.

2.1 Poisson

The Poisson model problem is a prototypical partial differential equation with many applications in physics. It involves the gradient ∇ and the divergence div. These differential operators read, for sufficiently smooth functions $v : \Omega \to \mathbb{R}$ and $q : \Omega \to \mathbb{R}^d$ with bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ (see [McL00, p. 89] for the definition of Lipschitz domain), dimension $d \in \mathbb{N}$, and $q = (q_n)_{n=1}^d$,

$$\nabla v = \begin{pmatrix} \partial v/\partial x_1 \\ \partial v/\partial x_2 \\ \vdots \\ \partial v/\partial x_d \end{pmatrix} \qquad \text{and} \qquad \operatorname{div} q = \sum_{n=1}^{d} \partial q_n/\partial x_n. \tag{2.1}$$

Given a function $f \in L^2(\Omega) := \{g : \Omega \to \mathbb{R} \mid \int_\Omega g^2 \,\mathrm{d}x < \infty\}$ in the Lebesgue space of square integrable functions, the Poisson model problem with homogeneous Dirichlet boundary conditions on the boundary $\partial\Omega$ of the domain Ω seeks the solution $u : \Omega \to \mathbb{R}$ and $p : \Omega \to \mathbb{R}^d$ to the first-order problem

$$-\operatorname{div} p = f \text{ in } \Omega, \qquad \nabla u - p = 0 \text{ in } \Omega, \qquad \text{and} \qquad u = 0 \text{ on } \partial\Omega. \tag{2.2}$$

The definition of the Laplace operator $\Delta := \operatorname{div} \nabla$ and the application of the identity $\nabla u = p$ to the first equation in (2.2) lead to the equivalent second-order problem

$$-\Delta u = f \text{ in } \Omega \qquad \text{and} \qquad u = 0 \text{ on } \partial\Omega. \tag{2.3}$$

Since the Poisson model problem (2.3) does not admit a strong solution in general (see [HL11, p. 65] for a counterexample), the well-posedness of (2.3) requires a generalization of the differential operators. The generalization bases on the concept of weak partial derivatives [Eva10, Sec. 5.2.1]. Define the space of infinitely differentiable functions with compact support $C_c^\infty(\Omega;\mathbb{R}^k) := \{\xi : \Omega \to \mathbb{R}^k \mid \xi \text{ is infinitely differentiable and has compact support}\}$ and the vector-valued Lebesgue space $L^2(\Omega;\mathbb{R}^k) := \{(q_n)_{n=1}^k \mid q_n \in L^2(\Omega)\}$ with inner product $(\bullet,\bullet)_{L^2(\Omega)} := \int_\Omega \bullet \cdot \bullet \,\mathrm{d}x$ and induced norm $\|\bullet\|_{L^2(\Omega)}$ for all $k \in \mathbb{N}$.

Definition 2.1.1 (Weak gradient and divergence)**.** *The function $\nabla v \in L^2(\Omega;\mathbb{R}^d)$ is called (weak) gradient of v and the function $\operatorname{div} q \in L^2(\Omega)$ is called (weak) divergence of $q \in L^2(\Omega;\mathbb{R}^d)$ if and only if*

$$(v, \operatorname{div}\xi)_{L^2(\Omega)} = -(\nabla v, \xi)_{L^2(\Omega)} \qquad \text{for all } \xi \in C_c^\infty(\Omega;\mathbb{R}^d), \tag{2.4a}$$
$$(\operatorname{div} q, \vartheta)_{L^2(\Omega)} = -(q, \nabla\vartheta)_{L^2(\Omega)} \qquad \text{for all } \vartheta \in C_c^\infty(\Omega;\mathbb{R}). \tag{2.4b}$$

The weak differential operators are unique and equal the strong operators in (2.1) for smooth functions. The spaces of functions with weak gradient and divergence read

$$H^1(\Omega) := \{v \in L^2(\Omega) \mid \nabla v \in L^2(\Omega;\mathbb{R}^d)\},$$
$$H(\operatorname{div},\Omega) := \{q \in L^2(\Omega;\mathbb{R}^d) \mid \operatorname{div} q \in L^2(\Omega)\}.$$

They are Hilbert spaces with norms [BS02, Thm. 1.3.2],[Mon03, Thm. 3.22]

$$\|\bullet\|_{H^1(\Omega)} := (\|\bullet\|^2_{L^2(\Omega)} + \|\nabla\bullet\|^2_{L^2(\Omega)})^{1/2} \quad \text{and} \quad \|\bullet\|_{H(\operatorname{div},\Omega)} := (\|\bullet\|^2_{L^2(\Omega)} + \|\operatorname{div}\bullet\|^2_{L^2(\Omega)})^{1/2}.$$

Set the closure $H^1_0(\Omega)$ of $C_c^\infty(\Omega;\mathbb{R})$ and $H_0(\operatorname{div},\Omega)$ of $C_c^\infty(\Omega;\mathbb{R}^d)$ with respect to the norm in $H^1(\Omega)$ and $H(\operatorname{div},\Omega)$, that is

$$H^1_0(\Omega) := \overline{C_c^\infty(\Omega;\mathbb{R})}^{\|\bullet\|_{H^1(\Omega)}} \qquad \text{and} \qquad H_0(\operatorname{div},\Omega) := \overline{C_c^\infty(\Omega;\mathbb{R}^d)}^{\|\bullet\|_{H(\operatorname{div},\Omega)}}. \tag{2.5}$$

Since $v \in H^1(\Omega)$ and $v = 0$ on $\partial\Omega$ is equivalent to $v \in H^1_0(\Omega)$ (see Theorem 5.1.15), Definition 2.1.1 and (2.5) show that the weak formulation of (2.3) seeks $u \in H^1_0(\Omega)$ with

$$(\nabla u, \nabla v)_{L^2(\Omega)} = (f, v)_{L^2(\Omega)} \qquad \text{for all } v \in H^1_0(\Omega). \tag{2.6}$$

Lemma 2.1.2 (Friedrichs inequality)**.** *There exists a positive constant $C_F < \infty$ with*

$$\|v\|_{L^2(\Omega)} \le C_F \|\nabla v\|_{L^2(\Omega)} \qquad \text{for all } v \in H^1_0(\Omega).$$

Proof. This lemma is proven in [BS02, Prop. 5.3.3]. □

Theorem 2.1.3 (Well-posedness)**.** *Given a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ and a right-hand side $f \in L^2(\Omega)$, there exists a unique solution $u \in H^1_0(\Omega)$ to (2.6).*

Proof. Lemma 2.1.2 implies the equivalence of the norms $\|\bullet\|_{H^1(\Omega)}$ and $\|\nabla\bullet\|_{L^2(\Omega)}$ in $H^1_0(\Omega)$. Thus, $H^1_0(\Omega)$ is a Hilbert space with inner product $(\nabla\bullet, \nabla\bullet)_{L^2(\Omega)}$ and so the Riesz representation theorem [BS02, Thm. 2.4.2] proves well-posedness of (2.6). □

2.2 Helmholtz

The Helmholtz equation arises from the wave equation with time-harmonic solutions and is, among others, important for radar and sonar technologies, noise scattering, and seismology. Given a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ with dimension $d \in \mathbb{N}$, a frequency $\omega > 0$, and a source term $f \in L^2(\Omega)$, the Helmholtz equation seeks the solution $u : \Omega \to \mathbb{R}$ to

$$-\Delta u - \omega^2 u = f \text{ in } \Omega \qquad \text{and} \qquad u = 0 \text{ on } \partial\Omega. \tag{2.7}$$

The equivalent first-order system seeks $u : \Omega \to \mathbb{R}$ and $p : \Omega \to \mathbb{R}^d$ with

$$-\operatorname{div} p - \omega^2 u = f \text{ in } \Omega, \qquad \nabla u - p = 0 \text{ in } \Omega, \qquad \text{and} \qquad u = 0 \text{ on } \partial\Omega. \tag{2.8}$$

A modification of the counterexample for the Poisson model problem (set the right-hand side $f - \omega^2 u$ with f and u from the counterexample for the Poisson model problem in [HL11, p. 65]) proves that a strong solution to (2.7) does not always exist. This motivates the weak formulation of (2.7), which seeks the solution $u \in H^1_0(\Omega)$ to

$$(\nabla u, \nabla v)_{L^2(\Omega)} - \omega^2 (u, v)_{L^2(\Omega)} = (f, v)_{L^2(\Omega)} \qquad \text{for all } v \in H^1_0(\Omega). \tag{2.9}$$

The well-posedness of (2.9) depends on the frequency ω and the following eigenvalues.

Theorem 2.2.1 (Dirichlet eigenvalues of $-\Delta$)**.** *There exist countably many eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \dots$ with eigenfunctions $\phi_k \in H^1_0(\Omega) \setminus \{0\}$ of the Laplace operator $-\Delta$, i.e.,*

$$(\nabla \phi_k, \nabla v)_{L^2(\Omega)} = \lambda_k \, (\phi_k, v)_{L^2(\Omega)} \qquad \text{for all } v \in H^1_0(\Omega) \text{ and } k \in \mathbb{N}.$$

The eigenvalues $\lambda_k \to \infty$ as $k \to \infty$, the eigenfunctions are orthonormal in $L^2(\Omega)$, that is $(\phi_k, \phi_\ell)_{L^2(\Omega)} = \delta_{k\ell}$ for all $k, \ell \in \mathbb{N}$, and the linear hull $\operatorname{span}\{\phi_k \mid k \in \mathbb{N}\}$ is dense in $H^1_0(\Omega)$, that is

$$H^1_0(\Omega) = \overline{\operatorname{span}\{\phi_k \mid k \in \mathbb{N}\}}^{\|\nabla \bullet\|_{L^2(\Omega)}} = \overline{\operatorname{span}\{\phi_k \mid k \in \mathbb{N}\}}^{\|\bullet\|_{H^1(\Omega)}}.$$

Proof. This result can be found in [BBF13, p. 15]. □

Theorem 2.2.2 (Well-posedness)**.** *Given a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, a frequency $\omega > 0$, and a right-hand side $f \in L^2(\Omega)$, there exists a unique solution $u \in H^1_0(\Omega)$ to (2.9) if and only if $\omega^2 \notin \{\lambda_1, \lambda_2, \dots\}$ is not a Dirichlet eigenvalue of the Laplace operator.*

Proof. This theorem is proven in [Bar15, p. 101]. The proof verifies the assumptions of the following theorem. □

Theorem 2.2.3 (Babuška-Lax-Milgram)**.** *Let $(X, \|\bullet\|_X)$ and $(Y, \|\bullet\|_Y)$ be Hilbert spaces and $b : X \times Y \to \mathbb{R}$ a bilinear form with $\{y \in Y \mid b(x, y) = 0 \text{ for all } x \in X\} = \{0\}$ and*

$$0 < \beta := \inf_{x \in X \setminus \{0\}} \sup_{y \in Y \setminus \{0\}} \frac{b(x, y)}{\|x\|_X \|y\|_Y} \leq \|b\| := \sup_{x \in X \setminus \{0\}} \sup_{y \in Y \setminus \{0\}} \frac{b(x, y)}{\|x\|_X \|y\|_Y} < \infty. \tag{2.10}$$

Then there exists for any functional $F \in Y^*$ in the dual Y^* of Y a unique solution $u \in X$ to $b(u, y) = F(y)$ for all $y \in Y$. The solution satisfies $\beta \|u\|_X \leq \|F\|_{Y^*}$.

Proof. This theorem is proven in [Bab71, Thm. 2.1]. □

2.3 Elasticity

Linear elasticity models the deformation and internal stresses of solid objects due to a given load $f \in L^2(\Omega; \mathbb{R}^d)$ with bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$. It involves the Lebesgue space of square integrable matrix-valued functions

$$L^2(\Omega; \mathbb{R}^{d \times d}) := \{q : \Omega \to \mathbb{R}^{d \times d} \mid q = (q_{k\ell})_{k,\ell=1}^d \text{ with } q_{k\ell} \in L^2(\Omega) \text{ for all } k, \ell = 1, \dots, d\}$$

and the spaces of vector-valued $H^1(\Omega)$ functions and matrix-valued $H(\mathrm{div},\Omega)$ functions

$$\begin{aligned}
H^1(\Omega;\mathbb{R}^d) &:= \{v \in L^2(\Omega;\mathbb{R}^d) \mid v = (v_k)_{k=1}^d \text{ with } v_k \in H^1(\Omega) \text{ for all } k = 1,\dots,d\},\\
H^1_0(\Omega;\mathbb{R}^d) &:= \{v \in L^2(\Omega;\mathbb{R}^d) \mid v = (v_k)_{k=1}^d \text{ with } v_k \in H^1_0(\Omega) \text{ for all } k = 1,\dots,d\},\\
H(\mathrm{div},\Omega;\mathbb{R}^{d\times d}) &:= \{q \in L^2(\Omega;\mathbb{R}^{d\times d}) \mid q = (q_{k\ell})_{k,\ell=1}^d \text{ with } (q_{k\ell})_{k=1}^d \in H(\mathrm{div},\Omega)\\
&\qquad\qquad \text{for all } \ell = 1,\dots,d\}.
\end{aligned}$$

The differential operators ∇ and div from (2.1) (resp. the weak differential operators from (2.4)) apply componentwise to the vector and tensor spaces, that is, for all $v = (v_k)_{k=1}^d \in H^1(\Omega;\mathbb{R}^d)$ and $q = (q_{k\ell})_{k,\ell=1}^d \in H(\mathrm{div},\Omega;\mathbb{R}^{d\times d})$,

$$\nabla v = (\nabla v_1\ \nabla v_2\ \dots\ \nabla v_d) \qquad \text{and} \qquad \mathrm{div}\, q = \begin{pmatrix} \mathrm{div}\,(q_{k1})_{k=1}^d \\ \mathrm{div}\,(q_{k2})_{k=1}^d \\ \vdots \\ \mathrm{div}\,(q_{kd})_{k=1}^d \end{pmatrix}.$$

Define for all matrices $A \in \mathbb{R}^{d\times d}$ the transposed matrix $A^\top$ and set the infinitesimal strain tensor $\varepsilon(v) := (\nabla v + (\nabla v)^\top)/2$ [CD98, p. 188] for all vector-valued functions $v \in H^1(\Omega;\mathbb{R}^d)$. Given positive Lamé constants λ and μ and the identity matrix $I_{d\times d} \in \mathbb{R}^{d\times d}$, define for all matrices $A = (A_{k\ell})_{k,\ell=1}^d \in \mathbb{R}^{d\times d}$ the fourth-order elasticity tensor [SH98, Eq. 2.1.16]

$$\mathbb{C}A := 2\mu A + \lambda\,\mathrm{tr}(A) I_{d\times d} \qquad \text{with trace } \mathrm{tr}(A) := \textstyle\sum_{k=1}^d A_{kk}. \tag{2.11}$$

The second-order formulation of linear elasticity seeks the solution $u : \Omega \to \mathbb{R}^d$ to

$$-\mathrm{div}\,\mathbb{C}\varepsilon(u) = f \text{ in } \Omega \qquad \text{and} \qquad u = 0 \text{ on } \partial\Omega. \tag{2.12}$$

An equivalent first-order formulation seeks $u : \Omega \to \mathbb{R}^d$ and $\sigma : \Omega \to \mathbb{R}^{d\times d}$ with

$$-\mathrm{div}\,\sigma = f \text{ in } \Omega, \qquad \mathbb{C}\varepsilon(u) - \sigma = 0 \text{ in } \Omega, \qquad \text{and} \qquad u = 0 \text{ on } \partial\Omega. \tag{2.13}$$

Set the inner product $(p,q)_{L^2(\Omega)} := \int_\Omega p : q \,\mathrm{d}x := \sum_{k,\ell=1}^d \int_\Omega p_{k\ell} q_{k\ell}\,\mathrm{d}x$ for all $p = (p_{k\ell})_{k,\ell=1}^d$ and $q = (q_{k\ell})_{k,\ell=1}^d$ in $L^2(\Omega;\mathbb{R}^{d\times d})$ and define the induced norm $\|\bullet\|_{L^2(\Omega)}$. The weak formulation of (2.12) seeks the solution $u \in H^1_0(\Omega;\mathbb{R}^d)$ to the variational problem

$$(\mathbb{C}\varepsilon(u), \nabla v)_{L^2(\Omega)} = (f, v)_{L^2(\Omega)} \qquad \text{for all } v \in H^1_0(\Omega;\mathbb{R}^d). \tag{2.14}$$

Since $\mathbb{C}\varepsilon(u)$ is symmetric, it is orthogonal to the asymmetric part $\nabla v - \varepsilon(v)$ of ∇v with $v \in H^1_0(\Omega;\mathbb{R}^d)$ in the sense that $(\mathbb{C}\varepsilon(u), \nabla v - \varepsilon(v))_{L^2(\Omega)} = 0$. Thus, (2.14) is equivalent to

$$(\mathbb{C}\varepsilon(u), \varepsilon(v))_{L^2(\Omega)} = (f, v)_{L^2(\Omega)} \qquad \text{for all } v \in H^1_0(\Omega;\mathbb{R}^d). \tag{2.15}$$

Lemma 2.3.1 (Properties of the elasticity tensor $\mathbb{C}$)**.** *For all matrices $A, B \in \mathbb{R}^{d\times d}$ the operator $\mathbb{C}^{1/2}A := \sqrt{2\mu}A + d^{-1}(\sqrt{2\mu + d\lambda} - \sqrt{2\mu})\,\mathrm{tr}(A) I_{d\times d}$ satisfies*

$$\mathbb{C}A = \mathbb{C}^{1/2}\mathbb{C}^{1/2}A, \qquad \mathbb{C}^{1/2}A : B = A : \mathbb{C}^{1/2}B, \qquad \textit{and} \qquad 2\mu\, A : A \le \mathbb{C}^{1/2}A : \mathbb{C}^{1/2}A.$$

The compliance tensor $\mathbb{C}^{-1}$ has the form $\mathbb{C}^{-1}A = (2\mu)^{-1}A - \lambda^{-1}(2\mu(2\mu + d\lambda))\,\mathrm{tr}(A) I_{d\times d}$ for all matrices $A \in \mathbb{R}^{d\times d}$ and satisfies $\mathbb{C}^{-1}A\mathbb{C}A = \mathbb{C}A\mathbb{C}^{-1}A = I_{d\times d}$.

Proof. This theorem follows from simple calculations. □

Lemma 2.3.2 (Korn's first inequality)**.** *There exists a positive constant* $C_{\mathrm{Korn}} \leq \sqrt{2}$ *with*

$$\|\nabla v\|_{L^2(\Omega)} \leq C_{\mathrm{Korn}} \|\varepsilon(v)\|_{L^2(\Omega)} \qquad \text{for all } v \in H^1_0(\Omega;\mathbb{R}^d).$$

Proof. This lemma follows from integration by parts without any compactness arguments, see for example [Fri47, Sec. 2]. □

Theorem 2.3.3 (Well-posedness)**.** *Let* $\Omega \subset \mathbb{R}^d$ *be a bounded Lipschitz domain, let* λ *and* μ *be positive Lamé constants, and let* $f \in L^2(\Omega;\mathbb{R}^d)$ *be a given load. Then there exists a unique solution* $u \in H^1_0(\Omega;\mathbb{R}^d)$ *to the variational problem (2.15).*

Proof. Lemma 2.3.1 proves the equivalence of $\|\bullet\|_{L^2(\Omega)}$ and $\|\mathbb{C}^{1/2}\bullet\|_{L^2(\Omega)}$ in $L^2(\Omega;\mathbb{R}^{d\times d})$ and Lemma 2.3.2 proves the equivalence of $\|\nabla\bullet\|_{L^2(\Omega)}$ and $\|\varepsilon(\bullet)\|_{L^2(\Omega)}$ in $H^1_0(\Omega;\mathbb{R}^d)$. Thus, the norm $\|\mathbb{C}^{1/2}\varepsilon(\bullet)\|_{L^2(\Omega)}$ is equivalent to $\|\nabla\bullet\|_{L^2(\Omega)}$ in the Hilbert space $H^1_0(\Omega;\mathbb{R}^d)$ and so the Riesz representation theorem implies the well-posedness of (2.15). □

Theorem 2.3.4 (Dirichlet eigenvalues of $-\operatorname{div}\mathbb{C}\varepsilon(\bullet)$)**.** *There exist countably many eigenvalues* $0 < \lambda_1 \leq \lambda_2 \leq \dots$ *with eigenfunctions* $\phi_k \in H^1_0(\Omega;\mathbb{R}^d) \setminus \{0\}$ *of the Lamé operator* $-\operatorname{div}\mathbb{C}\varepsilon(\bullet)$*, that is*

$$(\mathbb{C}\varepsilon(\phi_k), \varepsilon(v))_{L^2(\Omega)} = \lambda_k \, (\phi_k, v)_{L^2(\Omega)} \qquad \text{for all } v \in H^1_0(\Omega;\mathbb{R}^d) \text{ and } k \in \mathbb{N}.$$

The eigenvalues $\lambda_k \to \infty$ *as* $k \to \infty$*, the eigenfunctions are orthonormal in* $L^2(\Omega;\mathbb{R}^d)$*, that is* $(\phi_k, \phi_\ell)_{L^2(\Omega)} = \delta_{k\ell}$ *for all* $k, \ell \in \mathbb{N}$*, and the linear hull* $\operatorname{span}\{\phi_k \mid k \in \mathbb{N}\}$ *is dense in* $H^1_0(\Omega;\mathbb{R}^d)$*, that is*

$$H^1_0(\Omega;\mathbb{R}^d) = \overline{\operatorname{span}\{\phi_k \mid k \in \mathbb{N}\}}^{\|\varepsilon(\bullet)\|_{L^2(\Omega)}}.$$

Proof. This theorem is proven in [CL91, p. 720]. □

2.4 Maxwell

The time-harmonic Maxwell equations describe electromagnetic phenomena in $d = 3$ space dimensions. They involve the rotation operator, which reads, for sufficiently smooth vector-valued functions $v : \Omega \to \mathbb{R}^3$ with $v = (v_k)_{k=1}^3$ and bounded Lipschitz domain $\Omega \subset \mathbb{R}^3$,

$$\operatorname{curl} v = \begin{pmatrix} \partial v_3/\partial x_2 - \partial v_2/\partial x_3 \\ \partial v_1/\partial x_3 - \partial v_3/\partial x_1 \\ \partial v_2/\partial x_1 - \partial v_1/\partial x_2 \end{pmatrix}. \tag{2.16}$$

Let $\nu : \partial\Omega \to \mathbb{R}^3$ denote the outer unit normal vector. Given a frequency $\omega > 0$ and a current density $f \in L^2(\Omega;\mathbb{R}^3)$, the second-order formulation of the time-harmonic Maxwell equations seeks $u : \Omega \to \mathbb{R}^3$ with

$$\operatorname{curl}\operatorname{curl} u - \omega^2 u = f \text{ in } \Omega \qquad \text{and} \qquad u \times \nu = 0 \text{ on } \partial\Omega. \tag{2.17}$$

The substitution of $\operatorname{curl} u$ by v results in the equivalent first-order problem

$$\operatorname{curl} v - \omega^2 u = f \text{ in } \Omega, \qquad v - \operatorname{curl} u = 0 \text{ in } \Omega, \qquad \text{and} \qquad u \times \nu = 0 \text{ on } \partial\Omega.$$

Recall the space $C^\infty_c(\Omega;\mathbb{R}^3)$ of infinitely differentiable functions $\xi : \Omega \to \mathbb{R}^3$ with compact support from Section 2.1.

Definition 2.4.1 (Weak rotation)**.** *A function* $\operatorname{curl} v \in L^2(\Omega;\mathbb{R}^3)$ *is called (weak) rotation of* $v \in L^2(\Omega;\mathbb{R}^3)$ *if and only if*

$$(v, \operatorname{curl}\xi)_{L^2(\Omega)} = -(\operatorname{curl} v, \xi)_{L^2(\Omega)} \qquad \text{for all } \xi \in C_c^\infty(\Omega;\mathbb{R}^3). \tag{2.18}$$

The space of all functions with weak rotation in $L^2(\Omega;\mathbb{R}^3)$ reads

$$H(\operatorname{curl},\Omega) := \{v \in L^2(\Omega;\mathbb{R}^3) \mid \operatorname{curl} v \in L^2(\Omega;\mathbb{R}^3)\}.$$

It is a Hilbert space with respect to the norm $\|\bullet\|_{H(\operatorname{curl},\Omega)} := (\|\bullet\|^2_{L^2(\Omega)} + \|\operatorname{curl}\bullet\|^2_{L^2(\Omega)})^{1/2}$ [Mon03, Thm. 3.26]. Set the closure $H_0(\operatorname{curl},\Omega)$ of $C_c^\infty(\Omega;\mathbb{R}^3)$, that is

$$H_0(\operatorname{curl},\Omega) := \overline{C_c^\infty(\Omega;\mathbb{R}^3)}^{\|\bullet\|_{H(\operatorname{curl},\Omega)}}. \tag{2.19}$$

Since $u \in H(\operatorname{curl},\Omega)$ and $\nu \times u = 0$ on $\partial\Omega$ if and only if $u \in H_0(\operatorname{curl},\Omega)$ [Mon03, Thm. 3.33], (2.17)–(2.19) result in the weak formulation: Seek $u \in H_0(\operatorname{curl},\Omega)$ with

$$(\operatorname{curl} u, \operatorname{curl} v)_{L^2(\Omega)} - \omega^2 (u,v)_{L^2(\Omega)} = (f,v)_{L^2(\Omega)} \qquad \text{for all } v \in H_0(\operatorname{curl},\Omega). \tag{2.20}$$

The well-posedness of (2.20) depends on the frequency ω and the following eigenvalues.

Theorem 2.4.2 (Eigenvalues of curl curl)**.** *There exist countably many eigenvalues* $0 < \lambda_1 \le \lambda_2 \le \dots$ *with eigenfunctions* $\phi_k \in H_0(\operatorname{curl},\Omega)\setminus\{0\}$ *of the* curl curl *operator, that is*

$$(\operatorname{curl}\phi_k, \operatorname{curl} v)_{L^2(\Omega)} = \lambda_k\,(\phi_k, v)_{L^2(\Omega)} \qquad \text{for all } v \in H_0(\operatorname{curl},\Omega) \text{ and } k \in \mathbb{N}.$$

The eigenvalues $\lambda_k \to \infty$ *as* $k \to \infty$*, the eigenfunctions are orthonormal in* $L^2(\Omega;\mathbb{R}^3)$*, that is* $(\phi_k,\phi_\ell)_{L^2(\Omega)} = \delta_{k\ell}$ *for all* $k,\ell \in \mathbb{N}$*, and the linear hull* $\operatorname{span}\{\phi_k \mid k \in \mathbb{N}\}$ *is dense in the orthogonal complement of the kernel* $H_0(\operatorname{curl}{=}0,\Omega) := \{v_0 \in H_0(\operatorname{curl},\Omega) \mid \operatorname{curl} v_0 = 0\}$*, that is* $(v_0,\phi_k)_{L^2(\Omega)} = 0$ *for all* $v_0 \in H_0(\operatorname{curl}{=}0,\Omega)$ *and* $k \in \mathbb{N}$ *as well as*

$$H_0(\operatorname{curl},\Omega) = H_0(\operatorname{curl}{=}0,\Omega) \oplus \overline{\operatorname{span}\{\phi_k \mid k\in\mathbb{N}\}}^{\|\bullet\|_{H(\operatorname{curl},\Omega)}}.$$

Proof. This theorem is proven in [Mon03, Thm. 4.18]. □

Theorem 2.4.3 (Well-posedness)**.** *Given a bounded Lipschitz domain* $\Omega \subset \mathbb{R}^3$*, a frequency* $\omega > 0$*, and a right-hand side* $f \in L^2(\Omega;\mathbb{R}^3)$*, there exists a unique solution to (2.20) if and only if* $\omega^2 \notin \{\lambda_1,\lambda_2,\dots\}$ *is not an eigenvalue of the* curl curl *operator.*

Proof. Many textbooks, as for example [Mon03, Cor. 4.19], prove this theorem with the Fredholm alternative. An alternative approach is a spectral decomposition of the ansatz space $H_0(\operatorname{curl},\Omega)$ as for example in [DV98, Sec. 3.3] (for the variational problem in (2.17) augmented by a Lagrange multiplier). The latter approach verifies the assumptions of Theorem 2.2.3. □

2.5 Stokes

The Stokes problem models the flow of incompressible Newtonian fluids. It involves the differential operators ∇ and div from Section 2.3, a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ with $2 \leq d \in \mathbb{N}$, and an external body force $f \in L^2(\Omega;\mathbb{R}^d)$. The Stokes problem seeks the velocity field $u : \Omega \to \mathbb{R}^d$ and the pressure $p : \Omega \to \mathbb{R}$ with

$$-\operatorname{div} \nabla u + \nabla p = f \text{ in } \Omega, \quad \operatorname{div} u = 0 \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega, \quad \text{and} \quad \int_\Omega p \,\mathrm{d}x = 0. \tag{2.21}$$

Define the trace $\operatorname{tr}(A) := \sum_{k=1}^d A_{kk}$ and the deviatoric tensor $\operatorname{dev} A := A - d^{-1}\operatorname{tr}(A) I_{d\times d}$ with identity matrix $I_{d\times d} \in \mathbb{R}^{d\times d}$ for all matrices $A = (A_{k\ell})_{k,\ell=1}^d \in \mathbb{R}^{d\times d}$. The identity $\nabla p = \operatorname{div}(pI_{d\times d})$, the definition $\sigma := \nabla u - pI_{d\times d}$, and $\operatorname{tr}(\nabla u) = \operatorname{div} u$ imply the equivalence of (2.21) and the pseudostress-velocity formulation

$$-\operatorname{div} \sigma = f \text{ in } \Omega, \quad \operatorname{dev} \sigma - \nabla u = 0 \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega, \quad \text{and} \quad \int_\Omega \operatorname{tr}(\sigma) \,\mathrm{d}x = 0. \tag{2.22}$$

The equivalent problem (2.22) was introduced in [CLW04] and enjoys huge scientific activity, see for example [BCM16, CGS13b, CTVW10, CW07, CW10, CWZ10, CKP11, FGM09, GMS10, GMS11]. The weak formulation of (2.22) seeks the velocity $u \in H^1_0(\Omega;\mathbb{R}^d)$ and the pseudostress $\sigma \in \Sigma := \{\tau \in H(\operatorname{div},\Omega;\mathbb{R}^{d\times d}) \mid \int_\Omega \operatorname{tr}(\tau)\,\mathrm{d}x = 0\}$ with

$$\begin{aligned} (\operatorname{dev}\sigma,\tau)_{L^2(\Omega)} + (\operatorname{div}\tau,u)_{L^2(\Omega)} &= 0 && \text{for all } \tau \in \Sigma, \\ (\operatorname{div}\sigma,v)_{L^2(\Omega)} &= -(f,v)_{L^2(\Omega)} && \text{for all } v \in L^2(\Omega;\mathbb{R}^d). \end{aligned} \tag{2.23}$$

Theorem 2.5.1 (tr-dev-div constant)**.** *There exists a constant $C_{\mathrm{tdd}} < \infty$ with*

$$\sup_{\tau\in\Sigma\setminus\{0\}} \frac{\|\operatorname{tr}(\tau)\|^2_{L^2(\Omega)}}{\|\operatorname{dev}\tau\|^2_{L^2(\Omega)} + \|\operatorname{div}\tau\|^2_{L^2(\Omega)}} =: C^2_{\mathrm{tdd}} < \infty. \tag{2.24}$$

Proof. The proof can be found in [BBF13, Prop. 9.1.1]. □

Theorem 2.5.2 (Ladyzhenskaya-Babuška-Brezzi)**.** *Set the space $L^2_0(\Omega) := \{q \in L^2(\Omega) \mid \int_\Omega q\,\mathrm{d}x = 0\}$. There exists a positive constant*

$$0 < C_{\mathrm{LBB}} := \inf_{q\in L^2_0(\Omega)\setminus\{0\}} \sup_{v\in H^1_0(\Omega;\mathbb{R}^d)\setminus\{0\}} \frac{(q,\operatorname{div} v)_{L^2(\Omega)}}{\|q\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)}}. \tag{2.25}$$

Proof. This theorem is proven for general Lipschitz domains Ω in [Bog79, Thm. 1]. □

The combination of the two previous theorems and the general theory for mixed variational formulations from [BBF13, Sec. 4.2] implies the following result.

Theorem 2.5.3 (Well-posedness)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain and $f \in L^2(\Omega;\mathbb{R}^d)$. There exists a unique solution $(u,\sigma) \in H^1_0(\Omega;\mathbb{R}^d) \times \Sigma$ to (2.23).*

Proof. A detailed proof can be found in [CTVW10, Thm. 2.3]. □

3 Least-squares finite element method

Systems of partial differential equations often lead to saddle-point like optimization problems. These problems cause theoretical and practical difficulties like unstable discretizations and indefinite algebraic problems. This motivates finite element schemes that circumvents these challenges. Most of these schemes utilize stabilization techniques for mixed methods or minimize a residual. The latter approach includes the least-squares finite element method, which approximates the solution $\mathfrak{u} \in X$ to a partial differential equation by the discrete minimizer $\mathfrak{u}_h \in X_h \subset X$ of a functional $LS(f;\bullet) = \|\mathfrak{u} - \bullet\|_a^2$. If the norm $\|\bullet\|_a$ is equivalent to the norm $\|\bullet\|_X$ in the Hilbert space X in the sense that there exist constants $0 < \alpha \leq \beta < \infty$ with

$$\alpha\|x\|_X^2 \leq \|x\|_a^2 \leq \beta\|x\|_X^2 \qquad \text{for all } x \in X, \tag{3.1}$$

this leads to an efficient and reliable numerical scheme with built-in error control

$$\alpha\|\mathfrak{u} - \mathfrak{u}_h\|_X^2 \leq \|\mathfrak{u} - \mathfrak{u}_h\|_a^2 = LS(f;\mathfrak{u}_h) \leq \beta\|\mathfrak{u} - \mathfrak{u}_h\|_X^2. \tag{3.2}$$

Section 3.1.1 introduces four hypotheses (**H1**)–(**H4**) which characterize the equivalence constants α and β as eigenvalues. This allows the computation of the equivalence constants and so (3.2) results in guaranteed error bounds. Furthermore, Section 3.1.1 shows that if the discrete space X_h depends on a parameter $h > 0$ (for example the maximal mesh-size of the underlying triangulation) and satisfies a density property (**D**), the error estimator $LS(f;\mathfrak{u}_h)$ is asymptotically exact, that is

$$LS(f;\mathfrak{u}_h)/\|\mathfrak{u} - \mathfrak{u}_h\|_X^2 \to 1 \qquad \text{as } h \to 0. \tag{3.3}$$

Moreover, the error $\|\mathfrak{u}_h - \mathfrak{u}_{\text{best}}\|_X$ between the discrete solution $\mathfrak{u}_h$ and the best approximation $\mathfrak{u}_{\text{best}} = \arg\min_{x_h \in X_h} \|\mathfrak{u} - x_h\|_X$ in X_h with respect to the norm $\|\bullet\|_X$ converges faster than the best approximation error, that is the asymptotic best approximation property

$$\|\mathfrak{u}_h - \mathfrak{u}_{\text{best}}\|_X/\|\mathfrak{u} - \mathfrak{u}_{\text{best}}\|_X \to 0 \qquad \text{as } h \to 0. \tag{3.4}$$

The asymptotic exactness result (3.3) indicates the overestimation of the error $\|\mathfrak{u}-\mathfrak{u}_h\|_X^2$ by the natural guaranteed upper bound $\|\mathfrak{u} - \mathfrak{u}_h\|_X^2 \leq \alpha^{-1} LS(f;\mathfrak{u}_h)$ from (3.2) with factor α^{-1} as $h \to 0$. Section 3.1.2 remedies the poor efficiency of the error control (3.2) by the combination of a priori knowledge of the continuous eigenspectrum with additional information on the discrete eigenspectrum. This leads to a (offline) computable improved reliability constant $C(X_h) \leq \alpha^{-1}$ with $\|\mathfrak{u} - \mathfrak{u}_h\|_X^2 \leq C(X_h)\, LS(f;\mathfrak{u}_h)$.

Section 3.2.1 validates the four hypotheses (**H1**)–(**H4**) for the Poisson model problem and Section 3.2.2 generalizes the results from Section 3.2.1 for the Helmholtz equation, linear elasticity, and the Maxwell equations. Numerical experiments at the end of these

two sections illustrate the asymptotic exactness for the model problems and show the significant improvement of the guaranteed upper bound with the computable improved reliability constant $C(X_h)$ from Section 3.1.2.

Section 3.2.3 reduces the ansatz space for the Stokes problem to verify the four hypotheses (**H1**)–(**H4**) with the techniques from Section 3.2.1–3.2.2. Numerical experiments at the end of Section 3.2.3 indicate that the reduction of the ansatz space is necessary for the asymptotic exactness result (3.3), that is the asymptotic exactness result (and so (**H1**)–(**H4**)) does not hold for the full ansatz space. A decomposition of the full ansatz space into the reduced ansatz space and the orthogonal complement leads to the equivalence constants α and β from (3.1) for the Stokes LSFEM with full ansatz space. The equivalence constants are related to the Ladyzhenskaya-Babuška-Brezzi (LBB) constant.

Remark 3.0.1. *The results from Section 3.1.1–3.2.2 base on the work [CS18], where C. Carstensen and J. Storn prove the asymptotic exact error estimation (3.3) and introduce the improved guaranteed error bound* $\|\mathfrak{u} - \mathfrak{u}_h\|_X^2 \leq C(X_h) LS(f; \mathfrak{u}_h)$. *Beside these results, this chapter contains several new contributions to the analysis of LSFEMs. Among others, it utilizes the abstract framework from [CS18] to prove the asymptotic best approximation property (3.4). Moreover, it validates* (**H1**)–(**H4**) *for the Stokes problem (with reduced ansatz space) and computes the equivalence constants for the Stokes problem (with full ansatz space).*

3.1 Analysis of LSFEM

3.1.1 Asymptotic exactness

This section proves the asymptotic exactness result (3.3) and best approximation property (3.4) for LSFEMs with the following abstract setting. Suppose X is a real Hilbert space with inner products $(\bullet, \bullet)_X$ and $a(\bullet, \bullet)$. Let $(\bullet, \bullet)_X$ induce the norm $\|\bullet\|_X$ and assume the following four hypotheses.

(**H1**) There are countably many pairwise disjoint positive numbers $\mu(0) = 1, \mu(1), \mu(2), \ldots$ with closed eigenspace $E(\mu(j))$ and

$$a(\psi_j, x) = \mu(j)\,(\psi_j, x)_X \qquad \text{for all } j \in \mathbb{N}_0, \psi_j \in E(\mu(j)), \text{ and } x \in X.$$

(**H2**) The eigenspaces have finite dimensions $\dim E(\mu(j)) \in \mathbb{N}$ for all $j \in \mathbb{N}$ (while the dimension $\dim E(\mu(0)) \in \mathbb{N}_0 \cup \{\infty\}$ might be infinity or zero).

(**H3**) The linear hull of all eigenspaces is dense in X with respect to $\|\bullet\|_X$, that is

$$X = \overline{\operatorname{span}\{E(\mu(j)) \mid j \in \mathbb{N}_0\}}^{\|\bullet\|_X}.$$

(**H4**) The only accumulation point of $(\mu(j))_{j\in\mathbb{N}_0}$ is $\mu(0) = 1 = \lim_{j\to\infty} \mu(j)$.

Remark 3.1.1 (Complex spaces)**.** *This thesis focuses in real Hilbert spaces. However, the results easily extend to complex Hilbert spaces.*

Lemma 3.1.2 (Orthogonal eigenspaces)**.** *Suppose* (**H1**)*. The eigenspaces are orthogonal in the sense that for all* $\psi_j \in E(\mu(j)), \psi_k \in E(\mu(k))$ *with* $j,k \in \mathbb{N}_0$ *and* $j \neq k$

$$(\psi_j, \psi_k)_X = 0 \qquad \text{and} \qquad a(\psi_j, \psi_k) = 0. \tag{3.5}$$

Proof. All $\psi_j \in E(\mu(j))$ and $\psi_k \in E(\mu(k))$ with $j,k \in \mathbb{N}_0$ and $j \neq k$ satisfy

$$\mu(j)\,(\psi_j, \psi_k)_X = a(\psi_j, \psi_k) = a(\psi_k, \psi_j) = \mu(k)\,(\psi_k, \psi_j)_X.$$

Since $\mu(j) \neq \mu(k)$, this proves $(\psi_j, \psi_k)_X = 0 = a(\psi_j, \psi_k)$. □

Theorem 3.1.3 (Equivalence of norms)**.** *Suppose* (**H1**)*,* (**H3**)*, and* (**H4**)*. Then the squared norm* $\|\bullet\|_a^2 := a(\bullet,\bullet)$ *is equivalent to* $\|\bullet\|_X^2 := (\bullet,\bullet)_X$ *with equivalence constants*

$$0 < \inf_{j\in\mathbb{N}_0} \mu(j) = \alpha := \inf_{x\in X\setminus\{0\}} \frac{\|x\|_a^2}{\|x\|_X^2} \leq \sup_{j\in\mathbb{N}_0} \mu(j) = \beta := \sup_{x\in X\setminus\{0\}} \frac{\|x\|_a^2}{\|x\|_X^2} < \infty. \tag{3.6}$$

Proof. Given $x \in X$, (**H3**) allows for the decomposition $x = \sum_{j\in\mathbb{N}_0} \psi_j$ with $\psi_j \in E(\mu(j))$ for all $j \in \mathbb{N}_0$. Thus, (**H1**) and the orthogonality of the eigenfunctions (3.5) imply

$$\|x\|_a^2 = \sum_{j\in\mathbb{N}_0} \|\psi_j\|_a^2 = \sum_{j\in\mathbb{N}_0} \mu(j)\|\psi_j\|_X^2 \leq \sup_{k\in\mathbb{N}_0} \mu(k) \sum_{j\in\mathbb{N}_0} \|\psi_j\|_X^2 = \sup_{k\in\mathbb{N}_0} \mu(k)\|x\|_X^2.$$

This shows $\beta \leq \sup_{j\in\mathbb{N}_0} \mu(j)$. Equality follows from the identity $\|\psi_j\|_a^2 = \mu(j)\|\psi_j\|_X^2$ for all $j \in \mathbb{N}$ and $\psi_j \in E(\mu(j)) \neq \{0\}$. An analogous procedure implies $\alpha = \inf_{j\in\mathbb{N}_0} \mu(j)$. Since $1 = \mu(0)$ is the only accumulation point of eigenvalues (**H4**) and all eigenvalues are positive (**H1**), it holds $0 < \inf_{j\in\mathbb{N}_0} \mu(j) \leq 1 \leq \sup_{j\in\mathbb{N}_0} \mu(j) < \infty$. □

The asymptotic exactness result (3.3) utilizes the hypotheses (**H1**)–(**H4**) and the following density property (**D**). Let the discrete space X_h depend on the parameter $h > 0$ (for example the maximal mesh-size of the underlying triangulation $\mathcal{T}$) and suppose

$$\text{(D)} \qquad \lim_{h\to 0} \min\{\|x - x_h\|_X \mid x_h \in X_h\} = 0 \qquad \text{for all } x \in X.$$

Theorem 3.1.4 (Asymptotic equality of norms)**.** *Suppose* (**H1**)–(**H4**) *and* (**D**)*. For all* $\varepsilon > 0$ *there exists some* $\delta > 0$ *such that, for all* $h \in (0,\delta]$ *and* $x \in X_h^\perp := \{x \in X \mid a(x, x_h) = 0$ *for all* $x_h \in X_h\}$*,*

$$(1-\varepsilon)\|x\|_X^2 \leq \|x\|_a^2 \leq (1+\varepsilon)\|x\|_X^2. \tag{3.7}$$

Proof. Let $0 < \varepsilon < 1$ and recall $\mu(0), \mu(1), \ldots$ and $E(\mu(0)), E(\mu(1)), \ldots$ from (**H1**).
Step 1 (Decomposition of X). Define the index set $J(\varepsilon) := \{j \in \mathbb{N} \mid \varepsilon < |1 - \mu(j)|\}$ with complement $J^c(\varepsilon) := \mathbb{N}_0 \setminus J(\varepsilon)$. It is a consequence of (**H2**) and (**H4**) that the index set $J(\varepsilon)$ is finite. Set the finite dimensional space $X(J(\varepsilon)) := \operatorname{span}\{E(\mu(j)) \mid j \in J(\varepsilon)\}$ and the closure with respect to the norm $\|\bullet\|_X$

$$X(J^c(\varepsilon)) := \overline{\operatorname{span}\{E(\mu(j)) \mid j \in J^c(\varepsilon)\}}^{\|\bullet\|_X}.$$

Let $(\psi_1, \dots, \psi_m)$ be a orthonormal basis of $X(J(\varepsilon)) = \operatorname{span}\{\psi_1, \dots, \psi_m\}$ with respect to the inner product $(\bullet, \bullet)_X$. The density (**D**) leads to a parameter $h_0 > 0$ such that for all $h \le h_0$ and $k = 1, \dots, m$ there exists a $\psi_{h,k} \in X_h$ with

$$\|\psi_k - \psi_{h,k}\|_X \le m^{-1/2}\varepsilon. \tag{3.8}$$

Step 2 (Decomposition of $x \in X_h^\perp$). Let $x \in X_h^\perp$ with $h \le h_0$ and $\|x\|_X = 1$. The density (**H3**) leads to the decomposition $x = y + z$ with $y \in X(J(\varepsilon))$ and $z \in X(J^c(\varepsilon))$. The orthogonality of the eigenspaces from Lemma 3.1.2 and the Pythagorean theorem imply

$$1 = \|x\|_X^2 = \|y\|_X^2 + \|z\|_X^2 \qquad \text{and} \qquad \|x\|_a^2 = \|y\|_a^2 + \|z\|_a^2. \tag{3.9}$$

Step 3 (Upper bound for $\|y\|_a$). There exist coefficients $y_1, \dots, y_m \in \mathbb{R}$ with $y = \sum_{k=1}^m y_k \psi_k$. The triangle inequality, the upper bound (3.8), the Cauchy-Schwarz inequality in $\mathbb{R}^m$, and the orthonormality of $\psi_1, \dots, \psi_m$ imply for $y_h := \sum_{k=1}^m y_k \psi_{h,k} \in X_h$ that

$$\|y - y_h\|_X \le \sum_{k=1}^m |y_k| \, \|\psi_k - \psi_{h,k}\|_X \le m^{-1/2}\varepsilon \sum_{k=1}^m |y_k| \le \varepsilon (\sum_{k=1}^m y_k^2)^{1/2} = \varepsilon \|y\|_X.$$

This inequality, the identities $a(y, z) = 0 = a(y_h, x)$, the Cauchy-Schwarz inequality, and the equivalence of norms (3.6) prove

$$\|y\|_a^2 = a(y, x) = a(y - y_h, x) \le \beta \|y - y_h\|_X \le \varepsilon\beta \|y\|_X \le \varepsilon\alpha^{-1/2}\beta \|y\|_a. \tag{3.10}$$

Step 4 (Bounds for $\|z\|_a^2$). The function $z \in X(J^c(\varepsilon))$ decomposes into $z = \sum_{j \in J^c(\varepsilon)} z_j$ with $z_j \in E(\mu(j))$ for all $j \in J^c(\varepsilon)$. The density of $\operatorname{span}\{\psi_j \mid j \in J^c(\varepsilon)\}$ in $X(J^c(\varepsilon))$ with respect to $\|\bullet\|_X$ and $\|\bullet\|_a$ (due to the equivalence of norms (3.6)) and the orthogonality of the eigenspaces from Lemma 3.1.2 imply $\|z\|_X^2 = \sum_{j \in J^c(\varepsilon)} \|z_j\|_X^2$ and $\|z\|_a^2 = \sum_{j \in J^c(\varepsilon)} \|z_j\|_a^2$. The properties $1 - \varepsilon \le \mu(j) \le 1 + \varepsilon$ and $\|z_j\|_a^2 = \mu(j)\|z_j\|_X^2$ for all $j \in J^c(\varepsilon)$ result in

$$(1 - \varepsilon)\|z\|_X^2 \le \sum_{j \in J^c(\varepsilon)} \mu(j)\|z_j\|_X^2 = \sum_{j \in J^c(\varepsilon)} \|z_j\|_a^2 = \|z\|_a^2 \le (1 + \varepsilon)\|z\|_X^2. \tag{3.11}$$

Step 5 (Upper bound for $\|x\|_a$). The combination of (3.9)–(3.11) proves

$$\|x\|_a^2 = \|y\|_a^2 + \|z\|_a^2 \le \|y\|_a^2 + (1 + \varepsilon)\|z\|_X^2 \le 1 + \varepsilon + \varepsilon^2\beta^2/\alpha.$$

Step 6 (Lower bound for $\|x\|_a$). The equivalence of the norms (3.6) and the equations in (3.9)–(3.10) show $1 - \varepsilon^2\beta^2/\alpha^2 \le 1 - \|y\|_a^2/\alpha \le 1 - \|y\|_X^2 = \|z\|_X^2$. Consequently,

$$(1 - \varepsilon)(1 - \varepsilon^2\beta^2/\alpha^2) \le (1 - \varepsilon)\|z\|_X^2 \le \|z\|_a^2 \le \|y\|_a^2 + \|z\|_a^2 = \|x\|_a^2.$$

Relabelling ε and h_0 for sufficiently small ε concludes the proof. □

Lemma 3.1.5 (Characterization of best approximations)**.** *Let $X_h \subset X$ be a discrete subspace of X. The best approximation $x_h = \arg\min_{y_h \in X_h} \|x - y_h\|_a$ of an element $x \in X$ equals the unique solution to $a(x_h, y_h) = a(x, y_h)$ for all $y_h \in X_h$.*

Proof. This textbook result is proven in [BS02, Prop. 2.3.1]. □

LSFEMs approximate the solution $\mathfrak{u} \in X$ to a partial differential equation the minimization of the least-squares functional

$$\mathfrak{u}_h = \operatorname*{arg\,min}_{x_h \in X_h} LS(f; x_h) \qquad \text{with} \qquad LS(f; x_h) := \|\mathfrak{u} - x_h\|_a^2. \tag{3.12}$$

Lemma 3.1.5 characterizes the minimizer $\mathfrak{u}_h \in X_h$ as solution to the variational problem

$$a(\mathfrak{u}_h, x_h) = a(\mathfrak{u}, x_h) \qquad \text{for all } x_h \in X_h. \tag{3.13}$$

This implies the Galerkin orthogonality $\mathfrak{u} - \mathfrak{u}_h \in X_h^\perp := \{y \in X \mid a(y, x_h) = 0 \text{ for all } x_h \in X_h\}$ and so Theorem 3.1.4 shows the asymptotic exactness, that is, for all $h > 0$ exist constants $c(h) \nearrow 1$ and $C(h) \searrow 1$ with

$$c(h)\|\mathfrak{u} - \mathfrak{u}_h\|_X^2 \le LS(f; \mathfrak{u}_h) \le C(h)\|\mathfrak{u} - \mathfrak{u}_h\|_X^2. \tag{3.14}$$

An alternative notation for (3.14) is $\lim_{h\to 0} LS(f; \mathfrak{u}_h)/\|\mathfrak{u} - \mathfrak{u}_h\|_X^2 = 1$, but this ignores that $\|\mathfrak{u} - \mathfrak{u}_h\|_X^2$ (and so $LS(f; \mathfrak{u}_h)$) could equal zero for all $h \in (0, h_0)$ and some $h_0 > 0$. The following notation includes this special case.

Definition 3.1.6 ($\lim'_{h\to 0}$)**.** *Let $(a_h)_{h>0}, (b_h)_{h>0} \subset \mathbb{R}$ be sequences and $r \in \mathbb{R}$. Define*

$$\lim_{h\to 0}' a_h/b_h = r$$

if and only if there exist $c(h), C(h) \in \mathbb{R}$ with $c(h) \nearrow r$, $C(h) \searrow r$ as $h \to 0$ and

$$c(h)b_h \le a_h \le C(h)b_h \qquad \text{for all } h > 0. \tag{3.15}$$

Theorem 3.1.7 (Asymptotic exactness of the LS residual)**.** *Let $\mathfrak{u}_h \in X_h$ be the discrete minimizer in (3.12). Then* (**H1**)–(**H4**) *and* (**D**) *imply the asymptotic exactness result*

$$\lim_{h\to 0}' LS(f; \mathfrak{u}_h)/\|\mathfrak{u} - \mathfrak{u}_h\|_X^2 = 1. \tag{3.16}$$

Proof. Equation (3.14) and the Galerkin orthogonality $\mathfrak{u} - \mathfrak{u}_h \in X_h^\perp$ prove this theorem. □

This section concludes with a modification of Theorem 3.1.7. The modification shows that the discrete minimizer $\mathfrak{u}_h \in X_h$ with (3.12) is almost the best approximation in X_h of $\mathfrak{u}$ with respect to the norm $\|\bullet\|_X$ for sufficiently fine parameters $h > 0$.

Theorem 3.1.8 (Asymptotic best approximation property)**.** *Suppose* (**H1**)–(**H4**) *and* (**D**)*. The solution $\mathfrak{u}_h \in X_h$ to (3.12) and the best approximation $\mathfrak{u}_{\text{best}} := \arg\min_{x_h \in X_h} \|\mathfrak{u} - x_h\|_X$ of the exact solution $\mathfrak{u} \in X$ satisfy*

$$\lim_{h\to 0}' \|\mathfrak{u}_h - \mathfrak{u}_{\text{best}}\|_X / \|\mathfrak{u} - \mathfrak{u}_{\text{best}}\|_X = 0.$$

Proof. Recall the eigenvalues $\mu(j)$ and eigenspaces $E(\mu(j))$ from (**H1**). The pairwise disjunct positive numbers $\tilde{\mu}(j) := \mu(j)^{-1}$ and spaces $\tilde{E}(\tilde{\mu}(j)) := E(\mu(j))$ satisfy

$$(\psi_j, x)_X = \tilde{\mu}(j)\, a(\psi_j, x) \qquad \text{for all } j \in \mathbb{N}_0, \psi_j \in \tilde{E}(\tilde{\mu}(j)), \text{ and } x \in X.$$

The equivalence of norms (3.6) shows

$$X = \overline{\operatorname{span}\{E(\mu(j)) : j \in \mathbb{N}_0\}}^{\|\bullet\|_X} = \overline{\operatorname{span}\{\tilde{E}(\tilde{\mu}(j)) : j \in \mathbb{N}_0\}}^{\|\bullet\|_a}.$$

Moreover, the dimension $\dim \tilde{E}(\tilde{\mu}(j)) = \dim E(\mu(j)) \in \mathbb{N}$ for all $j \in \mathbb{N}$ and the only accumulation point of $(\tilde{\mu}(j))_{j\in\mathbb{N}_0}$ is $\lim_{j\to\infty} \mu(j)^{-1} = 1$. Thus, Theorem 3.1.7 proves

$$\lim_{h\to 0}{}' \|\mathfrak{u} - \mathfrak{u}_{\text{best}}\|_X^2 / \|\mathfrak{u} - \mathfrak{u}_{\text{best}}\|_a^2 = 1. \tag{3.17}$$

Case 1. For all $h > 0$ with $\|\mathfrak{u} - \mathfrak{u}_{\text{best}}\|_X = 0$ the functions $\mathfrak{u}_h = \mathfrak{u} = \mathfrak{u}_{\text{best}}$. Thus, (3.15) holds with $c(h) = 0 = C(h)$.

Case 2. For all $h > 0$ with $\|\mathfrak{u} - \mathfrak{u}_{\text{best}}\|_X > 0$ define the numbers $\delta_1(h) := \|\mathfrak{u} - \mathfrak{u}_h\|_X^2 / \|\mathfrak{u} - \mathfrak{u}_h\|_a^2 - 1$ and $\delta_2(h) := \|\mathfrak{u} - \mathfrak{u}_{\text{best}}\|_a^2 / \|\mathfrak{u} - \mathfrak{u}_{\text{best}}\|_X^2 - 1$. The Pythagorean theorem $\|\mathfrak{u}_h - \mathfrak{u}_{\text{best}}\|_X^2 = \|\mathfrak{u} - \mathfrak{u}_h\|_X^2 - \|\mathfrak{u} - \mathfrak{u}_{\text{best}}\|_X^2$ and $\|\mathfrak{u} - \mathfrak{u}_h\|_a^2 / \|\mathfrak{u} - \mathfrak{u}_{\text{best}}\|_a^2 \le 1$ imply

$$\begin{aligned} 0 &\le \frac{\|\mathfrak{u}_h - \mathfrak{u}_{\text{best}}\|_X^2}{\|\mathfrak{u} - \mathfrak{u}_{\text{best}}\|_X^2} = \frac{\|\mathfrak{u} - \mathfrak{u}_h\|_X^2}{\|\mathfrak{u} - \mathfrak{u}_{\text{best}}\|_X^2} - 1 = \frac{\|\mathfrak{u} - \mathfrak{u}_h\|_a^2}{\|\mathfrak{u} - \mathfrak{u}_{\text{best}}\|_a^2} \frac{\|\mathfrak{u} - \mathfrak{u}_h\|_X^2}{\|\mathfrak{u} - \mathfrak{u}_h\|_a^2} \frac{\|\mathfrak{u} - \mathfrak{u}_{\text{best}}\|_a^2}{\|\mathfrak{u} - \mathfrak{u}_{\text{best}}\|_X^2} - 1 \\ &\le \frac{\|\mathfrak{u} - \mathfrak{u}_h\|_X^2}{\|\mathfrak{u} - \mathfrak{u}_h\|_a^2} \frac{\|\mathfrak{u} - \mathfrak{u}_{\text{best}}\|_a^2}{\|\mathfrak{u} - \mathfrak{u}_{\text{best}}\|_X^2} - 1 = \delta_1(h) + \delta_2(h) + \delta_1(h)\delta_2(h). \end{aligned} \tag{3.18}$$

The asymptotic exactness properties (3.16) and (3.17) prove that $\delta_1(h)$ and $\delta_2(h)$ tend to zero as h vanishes. This implies the existence of numbers $0 = c(h) \le C(h)$ with (3.15). □

Remark 3.1.9 (Generalizations)**.** *The proofs of Theorem 3.1.7–3.1.8 are still valid for h-dependent right-hand sides $f(h) \in L^2(\Omega)$ and exact solutions $\mathfrak{u}(h) \in X$. More precisely, under the assumptions of Theorem 3.1.7, the best approximations $\mathfrak{u}_h = \arg\min_{x_h\in X_h} \|\mathfrak{u}(h) - x_h\|_a$ and $\mathfrak{u}_{\text{best}}(h) = \arg\min_{x_h\in X_h} \|\mathfrak{u}(h) - x_h\|_X$ satisfy*

$$\begin{aligned} &\lim_{h\to 0}{}' \, LS(f(h), \mathfrak{u}_h) / \|\mathfrak{u} - \mathfrak{u}_h\|_X^2 = 1, \\ &\lim_{h\to 0}{}' \, \|\mathfrak{u}_h - \mathfrak{u}_{\text{best}}(h)\|_X / \|\mathfrak{u} - \mathfrak{u}_{\text{best}}(h)\|_X = 0. \end{aligned}$$

This observation allows an application of the result to the DPG method in Section 5.2.1.

Remark 3.1.10 (Asymptotic best approximation in Galerkin FEMs)**.** *Mikhlin proved the asymptotic best approximation property of solutions to Galerkin finite element methods for compactly perturbed elliptic problems [KVZ+72, Thm. 16.2]. An alternative proof can be found in [BHP17, Thm. 20]. However, to the author's knowledge, the asymptotic best approximation property with respect to the norm $\|\bullet\|_X$ in LSFEMs has been unknown.*

3.1.2 Guarenteed error bounds

This section improves the error estimate (3.2) for problems with (**H1**)–(**H4**). The point of departure is the observation that Theorem 3.1.4 and (**D**) imply

$$\alpha \le \alpha(X_h) := \inf_{x\in X_h^\perp\setminus\{0\}} \|x\|_a^2 / \|x\|_X^2 \to 1 \qquad \text{as } h \to 0. \tag{3.19}$$

The property $\mathfrak{u}-\mathfrak{u}_h \in X_h^\perp := \{x \in X \mid a(x,x_h)=0 \text{ for all } x_h \in X_h\}$ for the exact solution $\mathfrak{u} \in X$ and the discrete minimizer $\mathfrak{u}_h = \arg\min_{x_h \in X_h} \|\mathfrak{u}-x_h\|_a$ leads to the estimate

$$\|\mathfrak{u}-\mathfrak{u}_h\|_X^2 \leq \alpha(X_h)^{-1}\|\mathfrak{u}-\mathfrak{u}_h\|_a^2.$$

This estimate motivates the approximation of $\alpha(X_h)^{-1}$ from above by a computable constant $C(X_h)$. The constant $C(X_h)$ involves the smallest discrete eigenvalues $\mu_{h,1} \leq \cdots \leq \mu_{h,n}$ for a fixed $n \leq \dim X_h$ with orthonormal eigenfunctions $\psi_{h,1},\dots,\psi_{h,n} \in X_h$ in the sense that $(\psi_{h,j},\psi_{h,k})_X = \delta_{jk}$ and

$$a(\psi_{h,j},x_h) = \mu_{h,j}\,(\psi_{h,j},x_h)_X \qquad \text{for all } x_h \in X_h \text{ and } j,k=1,\dots,n. \tag{3.20}$$

Furthermore, let $0<\mu_1 \leq \cdots \leq \mu_{n+1}$ be the smallest exact eigenvalues with orthonormal eigenfunctions $\psi_1,\dots,\psi_{n+1} \in X$ in the sense that $(\psi_j,\psi_k)_X = \delta_{jk}$ and

$$a(\psi_j,x) = \mu_j\,(\psi_j,x)_X \qquad \text{for all } x \in X \text{ and } j,k=1,\dots,n+1. \tag{3.21}$$

A comparison of exact and discrete eigenvalue clusters in $\{\mu_1,\dots,\mu_n\}$ and $\{\mu_{h,1},\dots,\mu_{h,n}\}$ is the basic idea in the approximation of $\alpha(X_h)^{-1} \leq C(X_h)$. Recall the eigenvalues $\mu(1),\mu(2),\dots$ and eigenspaces $E(\mu(1)),E(\mu(2)),\dots$ from (**H2**). It holds $\{\mu_1,\dots,\mu_{n+1}\} \subset \{\mu(j) : j \in \mathbb{N}_0\}$. Set for any compact interval $[\alpha',\beta'] \subset \mathbb{R}$ the spaces

$$\begin{aligned} E(\alpha',\beta') &:= \overline{\operatorname{span}\{E(\mu(j)) \mid j \in \mathbb{N}_0 \text{ and } \alpha' \leq \mu(j) \leq \beta'\}}^{\|\bullet\|_X} \quad \text{and} \\ E_h(\alpha',\beta') &:= \operatorname{span}\{\psi_{h,j} \mid j=1,\dots,n \text{ and } \alpha' \leq \mu_{h,j} \leq \beta'\}. \end{aligned}$$

Assume (**H1**)–(**H4**) and the following hypothesis:

(**H5**) Let $[\alpha_1,\beta_1],\dots,[\alpha_m,\beta_m] \subset \mathbb{R}$ be pairwise disjoint compact intervals with $m \leq n$ and $0<\alpha_1 \leq \alpha \leq \beta_1 < \alpha_2 \leq \beta_2 < \cdots \leq \beta_m < \alpha_{m+1}$, which satisfy, for all $\ell=1,\dots,m$,

$$\dim E(\alpha_\ell,\beta_\ell) = \dim E_h(\alpha_\ell,\beta_\ell) \quad \text{and} \quad X = E(\alpha_1,\beta_\ell) \oplus E(\alpha_{\ell+1},\beta).$$

The intervals from (**H5**) lead to the constant

$$C(X_h) := \alpha_{m+1}^{-1}\left(1+\sum_{k=1}^{m}\alpha_{k+1}\frac{\alpha_{m+1}-\alpha_k}{\alpha_k\beta_k}\frac{\beta_k-\alpha_k}{\alpha_{k+1}-\alpha_k}\right). \tag{3.22}$$

Theorem 3.1.11 (Improved GUB)**.** *Suppose* (**H1**)–(**H5**)*, then* $C(X_h)$ *from (3.22) satisfies*

$$\|x\|_X^2 \leq C(X_h)\|x\|_a^2 \qquad \textit{for all } x \in X_h^\perp. \tag{3.23}$$

Remark 3.1.12 (Rate of convergence in (3.3))**.** *Suppose* (**H1**)–(**H5**) *and* $\alpha_\ell = \mu(\ell)$ *for all* $\ell=1,\dots,m+1$ *with the smallest pairwise distinct eigenvalues* $\mu(1),\mu(2),\dots,\mu(m+1)$ *in* (**H1**)*. Theorem 3.1.11 shows that a small eigenvalue error* $\delta := \max_{\ell=1,\dots,m}(\beta_\ell-\mu(\ell))$ *of the discrete space guarantees*

$$\alpha(X_h)^{-1} \leq C(X_h) = \mu(m+1)^{-1} + \mathcal{O}(\delta).$$

Suppose the eigenvalue error is of the form $\delta = \mathcal{O}(h_{\max}^s)$ *for some rate* $s > 0$*. This and* $\mu(m+1) \to 1$ *as* $m \to \infty$ *imply the asymptotic exactness (3.3) in the sense that there exists a constant* $C(m)$*, which depends in particular on* $m \in \mathbb{N}$*, with*

$$\|\mathfrak{u} - \mathfrak{u}_h\|_X^2 / \|\mathfrak{u} - \mathfrak{u}_h\|_a^2 \leq \alpha(X_h)^{-1} \leq C(X_h) \leq \mu(m+1)^{-1} + C(m) h_{\max}^s. \tag{3.24}$$

In other words, for all $\varepsilon > 0$ *exists a constant* $C(\varepsilon) > 0$ *which tends to infinity as* ε *tends to zero and*

$$\|\mathfrak{u} - \mathfrak{u}_h\|_X^2 / \|\mathfrak{u} - \mathfrak{u}_h\|_a^2 \leq 1 + \varepsilon + C(\varepsilon) h_{\max}^s. \tag{3.25}$$

Remark 3.1.13 (Rate of convergence in (3.4))**.** *Estimate (3.18) involves the numbers* $\delta_1(h) := \|\mathfrak{u} - \mathfrak{u}_h\|_X^2 / \|\mathfrak{u} - \mathfrak{u}_h\|_a^2 - 1$ *and* $\delta_2(h) := \|\mathfrak{u} - \mathfrak{u}_{\mathrm{best}}\|_a^2 / \|\mathfrak{u} - \mathfrak{u}_{\mathrm{best}}\|_X^2 - 1$ *and reads*

$$\|\mathfrak{u}_h - \mathfrak{u}_{\mathrm{best}}\|_X^2 / \|\mathfrak{u} - \mathfrak{u}_{\mathrm{best}}\|_X^2 \leq \delta_1(h) + \delta_2(h) + \delta_1(h)\delta_2(h). \tag{3.26}$$

The assumptions of Remark 3.1.12 imply the existence of a constant $C_1(\varepsilon) > 0$ *and a rate* $s_1 > 0$ *with* $\delta_1(h) \leq \varepsilon + C_1(\varepsilon) h_{\max}^{s_1}$ *for all* $\varepsilon > 0$*. Interchanging the inner products* $(\bullet, \bullet)_X$ *and* $a(\bullet, \bullet)$ *(as in the proof of Theorem 3.1.8) allows to apply Theorem 3.1.11 to bound the error* $\|\mathfrak{u} - \mathfrak{u}_{\mathrm{best}}\|_a$ *with the best approximation* $\mathfrak{u}_{\mathrm{best}} = \arg\min_{x_h \in X_h} \|\mathfrak{u} - x_h\|_X$*. Thus, Remark 3.1.12 applies and shows the existence of a constant* $C_2(\varepsilon)$ *and a rate* $s_2 > 0$ *with* $\delta_2(h) \leq \varepsilon + C_2(\varepsilon) h_{\max}^{s_2}$ *for all* $\varepsilon > 0$*. Set* $s := \min\{s_1, s_2\}$*. The combination of (3.26) and the bounds for* $\delta_1(h)$ *and* $\delta_2(h)$ *yields the existence of a constant* $C(\varepsilon) > 0$ *with*

$$\|\mathfrak{u}_h - \mathfrak{u}_{\mathrm{best}}\|_X^2 / \|\mathfrak{u} - \mathfrak{u}_{\mathrm{best}}\|_X^2 \leq \varepsilon + C(\varepsilon) h_{\max}^s \qquad \text{for all } \varepsilon > 0. \tag{3.27}$$

Proof of Theorem 3.1.11. Step 1 (Decomposition of $x \in X_h^\perp$*).* Given any $x \in X_h^\perp \setminus \{0\}$, (**H2**) and (**H5**) yield $X = E(\alpha_1, \beta_1) \oplus \cdots \oplus E(\alpha_m, \beta_m) \oplus E(\alpha_{m+1}, \beta)$ with β from (3.6). This implies the existence of $x_1, \ldots, x_{m+1} \in X$ with $x = \sum_{j=1}^{m+1} x_j$ and $x_j \in E(\alpha_j, \beta_j)$ for all $j = 1, \ldots, m+1$ (where $\beta_{m+1} := \beta$). The pairwise orthogonality of the eigenspaces (3.5) shows $a(x_j, x_k) = 0 = (x_j, x_k)_X$ for all $j, k = 1, \ldots, m+1$ with $j \neq k$.

Step 2 (Existence of $x_{h,j} \in E_h(\alpha_1, \beta_j)$ *with* $x_j - x_{h,j} \in E(\alpha_{j+1}, \beta)$*).* Let $j \in \{1, \ldots, m\}$ and $p = \dim E(\alpha_1, \beta_j)$, so that $\psi_1, \ldots, \psi_p \in X$ form a basis of $E(\alpha_1, \beta_j)$. Since $\dim E(\alpha_1, \beta_j) = \dim E_h(\alpha_1, \beta_j)$, there exists a basis $\psi_{h,1}, \ldots, \psi_{h,p} \in X_h$ of $E_h(\alpha_1, \beta_j)$. It holds $X_h \subset X = E(\alpha_1, \beta_j) \oplus E(\alpha_{j+1}, \beta)$. Consequently, there exists a matrix $B = (B_{k\ell})_{k,\ell=1,\ldots,p} \in \mathbb{R}^{p \times p}$ with

$$\psi_{h,k} - \sum_{\ell=1}^{p} B_{k\ell} \psi_\ell \in E(\alpha_{j+1}, \beta) \qquad \text{for all } k = 1, \ldots, p.$$

To prove that B is invertible, let $\xi = (\xi_1, \ldots, \xi_p) \in \mathbb{R}^p$ with $B\xi = 0$. In other words, $\sum_{k=1}^p \xi_k B_{k\ell} = 0$ for all $\ell = 1, \ldots, p$. Define

$$y_h := \xi_1 \psi_{h,1} + \cdots + \xi_p \psi_{h,p} \in E_h(\alpha_1, \beta_j).$$

If $\xi \neq 0$, the discrete function $y_h \in E_h(\alpha_1, \beta_j) \setminus \{0\}$ satisfies $a(y_h, y_h)/(y_h, y_h)_X \leq \beta_j$. Furthermore, since

$$\sum_{k=1}^{p} \xi_k \sum_{\ell=1}^{p} B_{k\ell} \psi_\ell = \sum_{\ell=1}^{p} (\sum_{k=1}^{p} \xi_k B_{k\ell}) \psi_\ell = 0,$$

it holds

$$y_h = y_h - \sum_{k=1}^{p} \xi_k \sum_{\ell=1}^{p} B_{k\ell}\psi_\ell = \sum_{k=1}^{p} \xi_k \Big(\psi_{h,k} - \sum_{\ell=1}^{p} B_{k\ell}\psi_\ell\Big) \in E(\alpha_{j+1}, \beta).$$

This implies $\alpha_{j+1} \leq a(y_h, y_h)/(y_h, y_h)_X$ and contradicts $\beta_j < \alpha_{j+1}$. Therefore, $y_h = 0$ and $(\xi_1, \ldots, \xi_p) = 0$. This proves that B is invertible. Thus, there exist coefficients $b_{\ell 1}, \ldots, b_{\ell p} \in \mathbb{R}$ for all $\ell = 1, \ldots, p$ with

$$\psi_\ell - \sum_{k=1}^{p} b_{\ell k}\psi_{h,k} \in E(\alpha_{j+1}, \beta).$$

This implies for $x_j \in \operatorname{span}\{\psi_1, \ldots, \psi_p\}$ the existence of $x_{h,j} \in \operatorname{span}\{\psi_{h,1}, \ldots, \psi_{h,p}\}$ with

$$x_j - x_{h,j} \in E(\alpha_{j+1}, \beta) \qquad \text{and} \qquad x_{h,j} \in E_h(\alpha_1, \beta_j). \tag{3.28}$$

Step 3 (Upper bound for $\|x_{h,j}\|_X^2$). The Pythagorean theorem and the orthogonality $E(\alpha_1, \beta_j) \perp_a E(\alpha_{j+1}, \beta)$ and $E(\alpha_1, \beta_j) \perp_X E(\alpha_{j+1}, \beta)$ imply for x_j from Step 1 and $x_{h,j}$ from (3.28) that $\|x_j\|_a^2 = \|x_{h,j}\|_a^2 - \|x_j - x_{h,j}\|_a^2$ and $\|x_j - x_{h,j}\|_X^2 = \|x_{h,j}\|_X^2 - \|x_j\|_X^2$. Since $x_{h,j} \in E_h(\alpha_1, \beta_j)$, it holds $\|x_{h,j}\|_a^2 \leq \beta_j \|x_{h,j}\|_X^2$. Moreover, $x_j - x_{h,j} \in E(\alpha_{j+1}, \beta)$ implies $\alpha_{j+1}\|x_j - x_{h,j}\|_X^2 \leq \|x_j - x_{h,j}\|_a^2$ and $x_j \in E(\alpha_j, \beta_j)$ induces $\alpha_j \|x_j\|_X^2 \leq \|x_j\|_a^2$. This shows

$$\alpha_j \|x_j\|_X^2 \leq \|x_j\|_a^2 = \|x_{h,j}\|_a^2 - \|x_j - x_{h,j}\|_a^2 \leq \beta_j \|x_{h,j}\|_X^2 - \alpha_{j+1}(\|x_{h,j}\|_X^2 - \|x_j\|_X^2).$$

Consequently,

$$\|x_{h,j}\|_X^2 \leq \frac{\alpha_{j+1} - \alpha_j}{\alpha_{j+1} - \beta_j} \|x_j\|_X^2. \tag{3.29}$$

Step 4 (Upper bound for $\|x_j\|_a^2$). Case 1. Let $x_j \neq 0$. This step utilizes $a(x_j, x_j - x_{h,j}) = 0 = a(x, x_{h,j})$ and the Cauchy-Schwarz inequality to deduce

$$\|x_j\|_a^2 = a(x, x_j) = a(x, x_j - x_{h,j}) = a(x - x_j, x_j - x_{h,j}) \leq \|x - x_j\|_a \|x_j - x_{h,j}\|_a.$$

The combination with the Pythagoras theorem, $\|x - x_j\|_a^2 = \|x\|_a^2 - \|x_j\|_a^2$ and $\|x_j - x_{h,j}\|_a^2 = \|x_{h,j}\|_a^2 - \|x_j\|_a^2$, leads to $\|x\|_a^2\|x_j\|_a^2 + \|x_j\|_a^2\|x_{h,j}\|_a^2 \leq \|x\|_a^2\|x_{h,j}\|_a^2$. Given $x_j \neq 0$, it follows from $x_j - x_{h,j} \in E(\alpha_{j+1}, \beta)$ and $E(\alpha_j, \beta_j) \cap E(\alpha_{j+1}, \beta) = \{0\}$ from (3.28) that $x_{h,j} \neq 0$. Consequently, the division of the previous estimate by $\|x\|_a^2\|x_j\|_a^2\|x_{h,j}\|_a^2 \neq 0$ results in

$$\|x\|_a^{-2} + \|x_{h,j}\|_a^{-2} \leq \|x_j\|_a^{-2}. \tag{3.30}$$

Since $x_{h,j} \in E_h(\alpha_j, \beta_j)$ and $x_j \in E(\alpha_j, \beta_j)$ fulfil $\beta_j^{-1}\|x_{h,j}\|_X^{-2} \leq \|x_{h,j}\|_a^{-2}$ and $\|x_j\|_a^{-2} \leq \alpha_j^{-1}\|x_j\|_X^{-2}$, (3.29) leads in (3.30) to

$$\|x_j\|_X^2 \leq \left(\frac{1}{\alpha_j} - \frac{\alpha_{j+1} - \beta_j}{\beta_j(\alpha_{j+1} - \alpha_j)}\right) \|x\|_a^2. \tag{3.31}$$

Case 2. Estimate (3.31) is trivial for $x_j = 0$.

Step 5 (Lower bound for $\|x\|_a^2$). The estimate $\alpha_j\|x_j\|_X^2 \leq \|x_j\|_a^2$ for all $j = 1, \dots, m+1$ and the pairwise orthogonality of $x_1, \dots, x_{m+1} \in X$ prove

$$\sum_{j=1}^{m} \alpha_j \|x_j\|_X^2 + \alpha_{m+1}(\|x\|_X^2 - \sum_{j=1}^{m} \|x_j\|_X^2) \leq \sum_{j=1}^{m+1} \|x_j\|_a^2 = \|x\|_a^2. \tag{3.32}$$

Since $\alpha_j - \alpha_{m+1} < 0$, the lower bound decreases monotonically in $\|x_j\|_X^2$ for each $j = 1, \dots, m$ and fixed $\|x\|_X^2$. Hence, the substitution of (3.31) into (3.32) leads to

$$\begin{aligned} \alpha_{m+1}\|x\|_X^2 &\leq \|x\|_a^2 + \sum_{j=1}^{m} (\alpha_{m+1} - \alpha_j) \left(\frac{1}{\alpha_j} - \frac{\alpha_{j+1} - \beta_j}{\beta_j(\alpha_{j+1} - \alpha_j)} \right) \|x\|_a^2 \\ &= \left(1 + \sum_{j=1}^{m} \alpha_{j+1} \frac{\alpha_{m+1} - \alpha_j}{\alpha_j \beta_j} \frac{\beta_j - \alpha_j}{\alpha_{j+1} - \alpha_j} \right) \|x\|_a^2. \end{aligned}$$

□

Numerical realization

The following three-stage algorithm allows for the computation of intervals with (**H5**). It requires some a priori knowledge on the exact and discrete eigenspectrum and leads to the reliability constant $C(X_h)$ from Theorem 3.1.11.

Stage 1. Compute $N+1$ lower bounds $0 < \mu_1^{\text{low}} \leq \cdots \leq \mu_{N+1}^{\text{low}}$ for the smallest continuous eigenvalues in (3.21), i.e, $\mu_j^{\text{low}} \leq \mu_j$ for $j = 1, \dots, N+1$. This computation is independent of the current triangulation and done offline.

Stage 2. Given a discretization $X_h \subset X$, compute upper bounds for the smallest discrete eigenvalues $0 < \mu_{h,1} \leq \cdots \leq \mu_{h,N}$ with linear independently eigenfunctions $\psi_{h,1}, \dots, \psi_{h,N} \in X_h \setminus \{0\}$ such that

$$a(\psi_{h,\ell}, w_h) = \mu_{h,\ell}\,(\psi_{h,\ell}, w_h)_X \qquad \text{for all } \ell = 1, \dots, N \text{ and } w_h \in X_h. \tag{3.33}$$

This leads to $\mu_{h,1}^{\text{up}} \leq \cdots \leq \mu_{h,N}^{\text{up}}$ with $\mu_{h,\ell} \leq \mu_{h,\ell}^{\text{up}}$ for all $\ell = 1, \dots, N$.

Stage 3. Given the lower and upper eigenvalue bounds from Stage 1 and 2, compute $C(X_h)$ for all $n = 0, \dots, N$ via the subsequent routine in Algorithm 1.

Algorithm 1: Computation of $C(X_h)$

Input: lower and upper eigenvalue bounds from Stage 1 and 2

```
1 set α_1 := μ_1^low and m := 0;
2 for k = 1, 2, ..., n do
3 |   if μ_{h,k}^up < μ_{k+1}^low then
4 |   |   set m := m + 1, β_m := μ_{h,k}^up, and α_{m+1} := μ_{k+1}^low;
5 |   end
6 end
7 apply the formula from (3.22);
```

Output: the minimum $C(X_h)$ of the values from line 7 for $n = 0, \dots, N$

Proposition 3.1.14 (Computable GUB)**.** *Suppose* (**H1**)–(**H4**)*, then the three-stage algorithm leads to the reliability constant $C(X_h)$ in Theorem 3.1.11.*

Proof. Step 1 (Proof of $0 < \alpha_1 \leq \alpha \leq \beta_1 < \alpha_2 \leq \beta_2 < \cdots \leq \beta_m < \alpha_{m+1}$). For all $j = 1, \ldots, n$, the Rayleigh-Ritz principle leads to

$$\mu_j^{\text{low}} \leq \mu_j = \min_{\substack{X_j \subset X \\ \dim X_j = j}} \max_{\substack{x \in X_j \\ \|x\|_X = 1}} a(x,x) \leq \mu_{h,j} = \min_{\substack{X_{h,j} \subset X_h \\ \dim X_{h,j} = j}} \max_{\substack{x_h \in X_{h,j} \\ \|x_h\|_X = 1}} a(x_h, x_h) \leq \mu_{h,j}^{\text{up}}. \tag{3.34}$$

This proves $0 < \alpha_1 = \mu_1^{\text{low}} \leq \mu_1 = \alpha \leq \mu_{h,1}^{\text{up}} = \beta_1$. Moreover, for all $\ell = 1, \ldots, m$ there exists a $k \in \{1, \ldots, n\}$ such that $\beta_\ell = \mu_{h,k}^{\text{up}} < \mu_{k+1}^{\text{low}} = \alpha_{\ell+1}$. If $\ell < m$, it also holds $\alpha_{\ell+1} = \mu_{k+1}^{\text{low}} \leq \mu_{h,k+1}^{\text{up}} \leq \beta_{\ell+1}$.

Step 2 (Proof of $\dim E(\alpha_\ell, \beta_\ell) = \dim E_h(\alpha_\ell, \beta_\ell)$). Given an interval $[\alpha_\ell, \beta_\ell]$ with $\ell = 1, \ldots, m$, let $\ell_1 \in \{1, \ldots, n\}$ be the smallest index with $\alpha_\ell = \mu_{\ell_1}^{\text{low}}$ and let $\ell_2 \in \{\ell_1, \ldots, n\}$ be the biggest index with $\beta_\ell = \mu_{h,\ell_2}^{\text{up}}$. Then

$$\alpha_\ell = \mu_{\ell_1}^{\text{low}} \leq \mu_{\ell_1} \leq \mu_{\ell_1+1} \leq \cdots \leq \mu_{\ell_2} \leq \mu_{h,\ell_2}^{\text{up}} = \beta_\ell$$

implies $\ell_2 - \ell_1 + 1 \leq \dim E(\alpha_\ell, \beta_\ell)$. If $\ell_2 - \ell_1 + 1 < \dim E(\alpha_\ell, \beta_\ell)$, there exists an eigenpair $(\mu, \psi) \in [\alpha_\ell, \beta_\ell] \times X \setminus \{0\}$ with $a(\psi, x) = \mu\,(\psi, x)_X$ for all $x \in X$ and $(\psi, \psi_k)_X = 0$ for all $k = 1, \ldots, n+1$. The eigenvalue μ is strictly smaller than μ_{ℓ_2+1}. This contradicts the assumption that $\mu_1, \ldots, \mu_{\ell_2+1}$ are the smallest eigenvalues. Therefore, $\dim E(\alpha_\ell, \beta_\ell) = \ell_2 - \ell_1 + 1$. Similar arguments lead to $\dim E_h(\alpha_\ell, \beta_\ell) = \ell_2 - \ell_1 + 1$.

Step 3 (Proof of $X = E(\alpha_1, \beta_\ell) \oplus E(\alpha_{\ell+1}, \beta)$). For all $\ell = 1, \ldots, m$ there exists $k \in \{1, \ldots, n\}$ with $\mu_k \leq \mu_{h,k}^{\text{up}} = \beta_\ell < \alpha_{\ell+1} = \mu_{k+1}^{\text{low}} \leq \mu_{k+1}$. Let $\psi_j \in E(\mu(j))$ with $j \in \mathbb{N}$. Since $\mu_1, \ldots, \mu_n$ are the smallest eigenvalues with (3.21), it holds either $\mu(j) \leq \mu_k$ or $\mu_{k+1} \leq \mu(j)$. This reveals $\psi_j \in E(\alpha_1, \beta_\ell)$ or $\psi_j \in E(\alpha_{\ell+1}, \beta)$. Therefore, any eigenfunction belongs to $E(\alpha_1, \beta_\ell) \oplus E(\alpha_{\ell+1}, \beta)$. The density of the linear hull of eigenfunctions in X from (**H2**) implies $X = E(\alpha_1, \beta_\ell) \oplus E(\alpha_{\ell+1}, \beta)$. □

Remark 3.1.15 (hp-refinements)**.** *The Rayleigh-Ritz principle (3.34) shows that the upper bounds for the smallest discrete eigenvalues in (3.33) are upper bounds for the smallest discrete eigenvalues for any discrete space $\hat{X}_h$ with $X_h \subset \hat{X}_h$. Thus, the estimate $\|\mathfrak{u} - \hat{\mathfrak{u}}_h\|_X^2 \leq C(X_h) LS(f; \hat{\mathfrak{u}}_h)$ holds for the solution $\hat{\mathfrak{u}}_h \in \hat{X}_h$ to the LSFEM with discrete space $\hat{X}_h$ with $X_h \subset \hat{X}_h$. This enables the computation of guaranteed error bounds $\|\mathfrak{u} - \hat{\mathfrak{u}}_h\|_X^2 \leq C(X_h) LS(f; \hat{\mathfrak{u}}_h)$ for any (adaptive) hp-refinement.*

3.2 Application of LSFEM

3.2.1 Poisson

This section validates (**H1**)–(**H4**) for the Poisson model problem. Thereby, it exemplifies several techniques which Section 3.2.2 generalizes for a wider class of problems. The validation of (**H1**)–(**H4**) implies the results from Section 3.1.1–3.1.2. This includes the asymptotic exactness (3.3) and the asymptotic best approximation (3.4). Even though

many mathematical papers like [BG05, CLMM94, JP93, PCL94] investigate LSFEMs for the Poisson model problem, these results seem to be unknown. More precisely, there exist (to the author's knowledge) no asymptotic exactness results with standard discretizations (the result in [CCKP15] is caused by an unbalanced discretization).

The LSFEM in this section approximates the solution $\mathfrak{u} = (u, p) \in X := H^1_0(\Omega) \times H(\mathrm{div}, \Omega)$ to (2.2) by the minimization of the least-squares functional

$$LS(f; v, q) := \|\nabla v - q\|^2_{L^2(\Omega)} + \|f + \mathrm{div}\, q\|^2_{L^2(\Omega)} \qquad \text{for all } (v, q) \in X$$

over a discrete subspace $X_h \subset X$. Since $\nabla u = p$ and $-\mathrm{div}\, p = f$, it holds $LS(f; v, q) = \|(u, p) - (v, q)\|^2_a$ with squared norm

$$\|(v, q)\|^2_a := LS(0; v, q) = \|\nabla v - q\|^2_{L^2(\Omega)} + \|\mathrm{div}\, q\|^2_{L^2(\Omega)} \qquad \text{for all } (v, q) \in X. \tag{3.35}$$

Let the inner product $a(\bullet, \bullet)$ induce the norm $\|\bullet\|_a$ and let the inner product $(\bullet, \bullet)_X$ induce the norm $\|\bullet\|_X$ with

$$\|(v, q)\|^2_X := \|\nabla v\|^2_{L^2(\Omega)} + \|q\|^2_{L^2(\Omega)} + \|\mathrm{div}\, q\|^2_{L^2(\Omega)} \qquad \text{for all } (v, q) \in X. \tag{3.36}$$

Define the space of divergence free functions $H(\mathrm{div}{=}0, \Omega) := \{q_0 \in H(\mathrm{div}, \Omega) \mid \mathrm{div}\, q_0 = 0\}$. For all $j \in \mathbb{N}$ recall the Dirichlet eigenpairs $(\lambda_j, \phi_j) \in \mathbb{R} \times H^1_0(\Omega)$ of the Laplace operator from Theorem 2.2.1 and set $\mu_0 = 1$ and $\psi_0 \in \{0\} \times H(\mathrm{div}{=}0, \Omega)$,

$$\begin{aligned} \mu_{2j-1} &:= 1 - (\lambda_j + 1)^{-1/2} \quad \text{and} \quad \psi_{2j-1} := ((1 + \lambda_j)^{1/2}\phi_j, \nabla\phi_j) \in X, \\ \mu_{2j} &:= 1 + (\lambda_j + 1)^{-1/2} \quad \text{and} \quad \psi_{2j} := ((1 + \lambda_j)^{1/2}\phi_j, -\nabla\phi_j) \in X. \end{aligned} \tag{3.37}$$

Theorem 3.2.1 (Eigenvalues of $a(\bullet, \bullet)$ and $(\bullet, \bullet)_X$)**.** *The pairs $(\mu_j, \psi_j) \in \mathbb{R} \times X$ from (3.37) solve the eigenvalue problem*

$$a(\psi_j, x) = \mu_j\, (\psi_j, x)_X \qquad \textit{for all } x \in X \textit{ and } j \in \mathbb{N}_0. \tag{3.38}$$

The proof of Theorem 3.2.1 utilizes the following well-known Helmholtz decomposition.

Lemma 3.2.2 (Helmholtz decomposition)**.** *Any $q \in L^2(\Omega; \mathbb{R}^d)$ decomposes into $q = \nabla\xi + q_0$ with unique functions $\xi \in H^1_0(\Omega)$ and $q_0 \in H(\mathrm{div}{=}0, \Omega)$. The decomposition is orthogonal in $L^2(\Omega; \mathbb{R}^d)$, that is $(\nabla\xi, q_0)_{L^2(\Omega)} = 0$.*

Proof. Let $q \in L^2(\Omega; \mathbb{R}^d)$. Since $(\nabla\bullet, \nabla\bullet)_{L^2(\Omega)}$ is an inner product in the Hilbert space $H^1_0(\Omega)$, the Riesz representation theorem yields the existence of a unique $\xi \in H^1_0(\Omega)$ with

$$(\nabla\xi, \nabla w)_{L^2(\Omega)} = (q, \nabla w)_{L^2(\Omega)} \qquad \text{for all } w \in H^1_0(\Omega).$$

Thus, $q_0 := q - \nabla\xi$ satisfies $(q_0, \nabla w)_{L^2(\Omega)} = (q, \nabla w)_{L^2(\Omega)} - (\nabla\xi, \nabla w)_{L^2(\Omega)} = 0$ for all $w \in H^1_0(\Omega)$. This shows $(\nabla\xi, q_0)_{L^2(\Omega)} = 0$ and the definition of the (weak) divergence in (2.4) implies $\mathrm{div}\, q_0 = 0$. □

Proof of Theorem 3.2.1. Step 1 (Decomposition of the inner products). Let $(v,q),(w,r)\in X$. The Helmholtz decomposition from Lemma 3.2.2 yields the existence of $\xi,\vartheta\in H^1_0(\Omega)$ and $q_0,r_0\in H(\mathrm{div}{=}0,\Omega)$ with $q=\nabla\xi+q_0$ and $r=\nabla\vartheta+r_0$. The density of the Dirichlet eigenfunctions of the Laplace operator in $H^1_0(\Omega)$ implies the existence of coefficients $v_j,w_j,\xi_j,\vartheta_j\in\mathbb{R}$ with

$$v=\sum_{j\in\mathbb{N}}v_j\phi_j,\qquad w=\sum_{j\in\mathbb{N}}w_j\phi_j,\qquad \xi=\sum_{j\in\mathbb{N}}\xi_j\phi_j,\qquad\text{and}\qquad \vartheta=\sum_{j\in\mathbb{N}}\vartheta_j\phi_j. \tag{3.39}$$

The orthogonality of the normed eigenfunctions ϕ_j and q_0,r_0 leads to the formal calculation

$$\begin{aligned}
&a(v,q;w,r)=(\nabla v-q,\nabla w-r)_{L^2(\Omega)}+(\mathrm{div}\,q,\mathrm{div}\,r)_{L^2(\Omega)}\\
&=(\sum_{j\in\mathbb{N}}(v_j-\xi_j)\nabla\phi_j,\sum_{k\in\mathbb{N}}(w_k-\vartheta_k)\nabla\phi_k)_{L^2(\Omega)}+(\sum_{j\in\mathbb{N}}\lambda_jq_j\phi_j,\sum_{k\in\mathbb{N}}\lambda_kr_k\phi_k)_{L^2(\Omega)}+(q_0,r_0)_{L^2(\Omega)}\\
&=\sum_{j\in\mathbb{N}}\begin{pmatrix}v_j\\ \xi_j\end{pmatrix}\cdot\begin{pmatrix}\lambda_j & -\lambda_j\\ -\lambda_j & \lambda_j+\lambda_j^2\end{pmatrix}\begin{pmatrix}w_j\\ \vartheta_j\end{pmatrix}+(q_0,r_0)_{L^2(\Omega)}.
\end{aligned}$$

Similar arguments result in

$$(v,q;w,r)_X=\sum_{j\in\mathbb{N}}\begin{pmatrix}v_j\\ \xi_j\end{pmatrix}\cdot\begin{pmatrix}\lambda_j & 0\\ 0 & \lambda_j+\lambda_j^2\end{pmatrix}\begin{pmatrix}w_j\\ \vartheta_j\end{pmatrix}+(q_0,r_0)_{L^2(\Omega)}.$$

Step 2 (Computation of eigenpairs). The decomposition of the inner products in Step 1 shows that $\mu_0=1$ and any element $\psi_0=(0,q_0)$ with $q_0\in H(\mathrm{div}{=}0,\Omega)$ satisfy (3.38). For all $j\in\mathbb{N}$ and $(w,r)\in X$ the decomposition in Step 1 results in

$$\begin{aligned}
a(\psi_{2j-1};w,r)&=\begin{pmatrix}(1+\lambda_j)^{1/2}\\ 1\end{pmatrix}\cdot\begin{pmatrix}\lambda_j & -\lambda_j\\ -\lambda_j & \lambda_j+\lambda_j^2\end{pmatrix}\begin{pmatrix}w_j\\ \vartheta_j\end{pmatrix}\\
&=\mu_{2j-1}\begin{pmatrix}(1+\lambda_j)^{1/2}\\ 1\end{pmatrix}\cdot\begin{pmatrix}\lambda_j & 0\\ 0 & \lambda_j+\lambda_j^2\end{pmatrix}\begin{pmatrix}w_j\\ \vartheta_j\end{pmatrix}=\mu_{2j-1}\,(\psi_{2j-1};w,r)_X.
\end{aligned}$$

Analogously, $a(\psi_{2j};w,r)=\mu_{2j}\,(\psi_{2j};w,r)_X$ follows for all $j\in\mathbb{N}$ and $(w,r)\in X$. □

Theorem 3.2.3 (Properties of the inner products)**.** *The inner products $a(\bullet,\bullet)$ and $(\bullet,\bullet)_X$ in $X=H^1_0(\Omega)\times H(\mathrm{div},\Omega)$, which induce the norms $\|\bullet\|_a$ and $\|\bullet\|_X$ from (3.35) and (3.36), satisfy* (**H1**)–(**H4**).

Proof of (**H1**). Let $(\mu_j,\psi_j)\in\mathbb{R}\times X$ from (3.37) for all $j\in\mathbb{N}$ and $\mu_0=1$. Set pairwise disjoint numbers $\mu(0)=1,\mu(1),\mu(2),\dots$ with $\{\mu(j)\mid j\in\mathbb{N}_0\}=\{\mu_j\mid j\in\mathbb{N}_0\}$. Define the eigenspaces $E(\mu(j)):=\mathrm{span}\{\psi_k\mid k\in\mathbb{N}\text{ and }\mu_k=\mu(j)\}$ for all $j\in\mathbb{N}$ and $E(\mu(0)):=\{0\}\times H(\mathrm{div}{=}0,\Omega)$. Then Theorem 3.2.1 implies (**H1**).

Proof of (**H4**). Since the Dirichlet eigenvalues of the Laplace operator $\lambda_j\to\infty$ as $j\to\infty$, the definition (3.37) implies $\mu_j\to1$ as $j\to\infty$. Thus, one is the only accumulation point of $\{\mu(j)\mid j\in\mathbb{N}_0\}=\{\mu_k\mid k\in\mathbb{N}_0\}$.

Proof of (**H2**). Since the eigenfunctions $\psi_1,\psi_2,\dots$ from (3.37) are linearly independent, the definition of $E(\mu(j))$ shows, with counting measure $|\bullet|$, that

$$\dim E(\mu(j))=|\{k\in\mathbb{N}\mid\mu_k=\mu(j)\}|\qquad\text{for all } j\in\mathbb{N}.$$

The proof of (**H4**) shows that one is the only accumulation point of $\{\mu_k \mid k \in \mathbb{N}_0\}$. Hence, $\mu(j) \neq 1$ for all $j \in \mathbb{N}$ implies $|\{k \in \mathbb{N} \mid \mu_k = \mu(j)\}| \in \mathbb{N}$.

Proof of (**H3**). Let $x = (v, q) \in X$ with decomposition $v = \sum_{j\in\mathbb{N}} v_j\phi_j$ and $q = q_0 + \sum_{j\in\mathbb{N}} q_j\nabla\phi_j$ as in (3.39). Recall $\psi_1, \psi_2, \ldots$ from (3.37) and set

$$x_k := (0, q_0) + \frac{1}{2}\sum_{j=1}^{k}\left(\frac{v_j}{(1+\lambda_j)^{1/2}} + q_j\right)\psi_{2j-1} + \left(\frac{v_j}{(1+\lambda_j)^{1/2}} - q_j\right)\psi_{2j} \quad \text{for all } k \in \mathbb{N}.$$

The definition of $E(\mu(j))$ in Step 1 reveals $x_k \in \operatorname{span}\{E(\mu(j)) \mid j \in \mathbb{N}_0\}$ for all $k \in \mathbb{N}$. A calculation shows $x - x_k = \sum_{j=k+1}^{\infty}(v_j\phi_j, q_j\nabla\phi_j)$ and so the orthogonality of the eigenfunctions results in

$$\begin{aligned}\|x - x_k\|_X^2 &= \sum_{j=k+1}^{\infty} v_j^2\|\nabla\phi_j\|_{L^2(\Omega)}^2 + \sum_{j=k+1}^{\infty} q_j^2\|\nabla\phi_j\|_{L^2(\Omega)}^2 + \sum_{j=k+1}^{\infty} q_j^2\|\Delta\phi_j\|_{L^2(\Omega)}^2 \\ &= \sum_{j=k+1}^{\infty} v_j^2\lambda_j + \sum_{j=k+1}^{\infty} q_j^2(\lambda_j + \lambda_j^2).\end{aligned}$$

The density of $\{\phi_j \mid j \in \mathbb{N}\}$ in $H_0^1(\Omega)$ with respect to $\|\nabla\bullet\|_{L^2(\Omega)}$ and $\operatorname{span}\{\nabla\phi_j \mid j \in \mathbb{N}\}$ in $H(\operatorname{div}, \Omega) \cap \nabla H_0^1(\Omega)$ with respect to $\|\bullet\|_{H(\operatorname{div},\Omega)}$ imply that the infinite series $\sum_{j=k+1}^{\infty} v_j^2\lambda_j$ and $\sum_{j=k+1}^{\infty} q_j^2(\lambda_j + \lambda_j^2)$ tend to zero as $k \to \infty$. Thus,

$$x \in \overline{\operatorname{span}\{E(\mu(j)) \mid j \in \mathbb{N}_0\}}^{\|\bullet\|_X}. \qquad \square$$

Numerical experiments

The remainder of this section presents numerical experiments which illustrate the asymptotic exactness and investigate the improved GUB from the three-stage algorithm in Section 3.1.2. The experiments base on the open source computing platform FEniCS. A detailed description of the implemented routines is postponed to Appendix A.1.

Let $\mathcal{T}$ be a regular triangulation of the bounded Lipschitz domain $\Omega \subset \mathbb{R}^2$ into triangles (see [Cia78, p. 38] for the definition of regular triangulation). Define the space $\mathbb{P}_k(T; \mathbb{R}^\ell)$ of polynomials $g : T \to \mathbb{R}^\ell$ of total degree at most $k \in \mathbb{N}_0$ for all $T \in \mathcal{T}$ and $\ell \in \mathbb{N}$. Let $RT_{k-1}(T) := \mathbb{P}_{k-1}(T; \mathbb{R}^d) + \mathbb{P}_k(T; \mathbb{R})\,\mathrm{id} \subset \mathbb{P}_k(T; \mathbb{R}^d)$ with the identity map $\mathrm{id} : T \to T$ for all $k \in \mathbb{N}$ and $T \in \mathcal{T}$. The Courant and Raviart-Thomas finite element spaces read

$$\begin{aligned}&S^k(\mathcal{T}) := \{v_h \in H^1(\Omega) \mid v_h|_T \in \mathbb{P}_k(T; \mathbb{R}) \text{ for all } T \in \mathcal{T}\}, \quad S_0^k(\mathcal{T}) := H_0^1(\Omega) \cap S^k(\mathcal{T}), \\ &RT_{k-1}(\mathcal{T}) := \{q_h \in H(\operatorname{div}, \Omega) \mid q_h|_T \in RT_{k-1}(T) \text{ for all } T \in \mathcal{T}\}. \end{aligned} \tag{3.40}$$

The discrete space $X_h := S_0^k(\mathcal{T}) \times RT_{k-1}(\mathcal{T})$ satisfies the density property (**D**) for all $k \in \mathbb{N}$ as the maximal mesh-size of the regularly refined triangulations $\mathcal{T}$ (that is the angles of the triangles $T \in \mathcal{T}$ do not degenerate) tends to zero [Bra07, Chap. 3.5], [Bar15, Lem. 3.6]. Since Theorem 3.2.3 implies (**H1**)–(**H4**), the density property (**D**) and Theorem 3.1.7–3.1.8 prove the asymptotic exactness of the least-squares residual $LS(f; \mathfrak{u}_h)$ and the asymptotic best approximation property of the discrete solution $\mathfrak{u}_h = \arg\min_{x_h \in X_h} LS(f; x_h)$ for all polynomial degrees $k \in \mathbb{N}$.

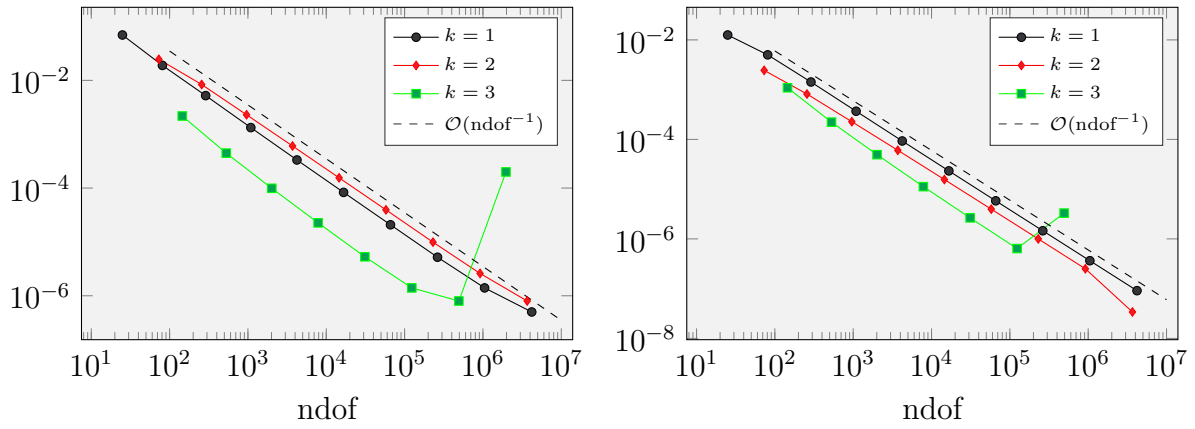


Figure 3.1: Distance $1 - LS(f;\mathfrak{u}_h)/\|\mathfrak{u}-\mathfrak{u}_h\|_X^2$ (left) and the ratio $\|\mathfrak{u}_h-\mathfrak{u}_{\text{best}}\|_X^2/\|\mathfrak{u}-\mathfrak{u}_{\text{best}}\|_X^2$ (right) with $k = 1, 2, 3$ in Experiment 1

Experiment 1 (Asymptotic exactness with known solution). The first experiment illustrates the asymptotic exactness and best approximation property. Let $\Omega = (0,1)^2$ be the unit square domain and define the right-hand side $f(x,y) = 2(x - x^2 + y - y^2)$ for all $(x,y) \in \Omega$. The solution reads $\mathfrak{u} = (u, \nabla u) = \arg\min_{x\in X} LS(f;x)$ with $u(x,y) = x(1-x)y(1-y)$ for all $(x,y) \in \Omega$. The experiment computes the discrete solution $\mathfrak{u}_h = \arg\min_{x_h\in X_h} LS(f;x_h)$ to the LSFEM and the best approximation $\mathfrak{u}_{\text{best}} = \arg\min_{x_h\in X_h}\|\mathfrak{u} - x_h\|_X$ with $X_h = S_0^k(\mathcal{T}) \times RT_{k-1}(\mathcal{T})$, uniformly refined triangulations $\mathcal{T}$ of Ω, and polynomial degrees $k = 1, 2, 3$. Figure 3.1 displays the distance $1 - LS(f;\mathfrak{u}_h)/\|\mathfrak{u} - \mathfrak{u}_h\|_X^2$, which is positive in all computations and vanishes as the maximal mesh-size $h_{\max} = \max\{\text{diam}(T) \mid T \in \mathcal{T}\}$ tends to zero. The experiment indicates

$$LS(f;\mathfrak{u}_h)/\|\mathfrak{u} - \mathfrak{u}_h\|_X^2 = 1 + \mathcal{O}(h_{\max}^2) = 1 + \mathcal{O}(\text{ndof}^{-1}) \qquad \text{for } k = 1, 2. \tag{3.41}$$

The rate of convergence is slightly better for $k = 3$. For more than than 10^5 degrees of freedom numerical difficulties occur. More precisely, an inexact computation of the system matrix (see Remark A.1.3) and the floating-point arithmetic cause the direct solver MUMPS to compute a solution $\tilde{\mathfrak{u}}_h \in X_h$ with tiny error $0 < \delta = \|\mathfrak{u}_h - \tilde{\mathfrak{u}}_h\|_a^2 \ll 1$. The Pythagorean theorem shows

$$\frac{\delta}{\|\mathfrak{u} - \tilde{\mathfrak{u}}_h\|_X^2} = \frac{LS(f;\tilde{\mathfrak{u}}_h)}{\|\mathfrak{u} - \tilde{\mathfrak{u}}_h\|_X^2} - \frac{LS(f;\mathfrak{u}_h)}{\|\mathfrak{u} - \tilde{\mathfrak{u}}_h\|_X^2}. \tag{3.42}$$

Since the error $\|\mathfrak{u}-\tilde{\mathfrak{u}}_h\|_X^2$ is very small on fine meshes (for example $\|\mathfrak{u}-\tilde{\mathfrak{u}}_h\|_X^2 = 4.63\times10^{-16}$ for $k = 3$ and $\text{ndof} = 493057$), a moderate numerical error δ causes a large ratio in (3.42). This indicates $0 \ll |LS(f;\mathfrak{u}_h)/\|\mathfrak{u} - \mathfrak{u}_h\|_X^2 - LS(f;\tilde{\mathfrak{u}}_h)/\|\mathfrak{u} - \tilde{\mathfrak{u}}_h\|_X^2|$ for fine meshes and explains the outliers.

If (3.41) holds for $k = 1, 2, 3$, the combination with (3.25) (and $\delta = \mathcal{O}(h_{\max}^{2k})$ due to the regularity of the smooth eigenfunctions) results for all $k = 1, 2, 3$ in the existence of constants $C(\varepsilon) > 0$ with

$$h_{\max}^2 \le \varepsilon + C(\varepsilon)h_{\max}^{2k} \qquad \text{for all } \varepsilon > 0 \text{ and } h_{\max} \in (0, 1]. \tag{3.43}$$

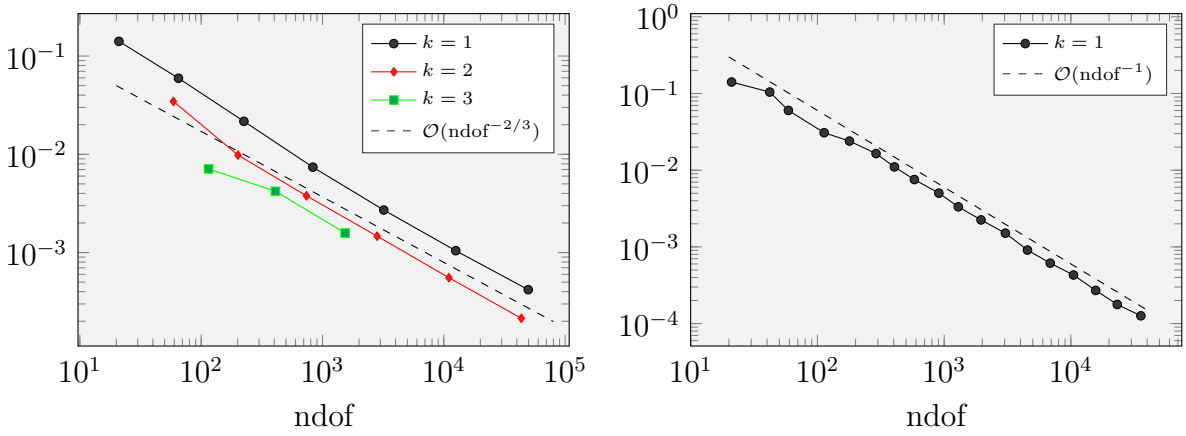


Figure 3.2: Distance $1 - LS(f;\mathfrak{u}_h)/\|\mathfrak{u}_{\mathrm{ref}} - \mathfrak{u}_h\|_X^2$ with uniformly (left) and adaptively (right) refined meshes for $k = 1, 2, 3$ in Experiment 2

Let $2\varepsilon = h_{\max}^2$ and $k = 1, 2, 3$, then (3.43) reads $1/2 \leq C(\varepsilon)(2\varepsilon)^{k-1}$ and indicates the lower bound

$$(2\varepsilon)^{1-k} \leq 2\, C(\varepsilon) \qquad \text{for all } \varepsilon > 0.$$

This suggests that $C(\varepsilon)$ in (3.25) behaves at best like a constant function for $k = 1$, like the inverse of an affine function for $k = 2$, and like the inverse of a quadratic function for $k = 3$ with respect to ε.

The right-hand side of Figure 3.1 displays the ratio $\|\mathfrak{u}_h - \mathfrak{u}_{\mathrm{best}}\|_X^2 / \|\mathfrak{u} - \mathfrak{u}_{\mathrm{best}}\|_X^2$. For $k = 1, 2$ the speed of convergence reads $\mathcal{O}(h_{\max}^2) = \mathcal{O}(\text{ndof}^{-1})$, for $k = 3$ the rate seems to be slightly better. As discussed before, numerical difficulties cause the outliers for $k = 2$, ndof $= 3674113$ and $k = 3$, ndof $= 493057$. The experiment and (3.18) suggest that the ratio $\|\mathfrak{u}_h - \mathfrak{u}_{\mathrm{best}}\|_X^2 / \|\mathfrak{u} - \mathfrak{u}_{\mathrm{best}}\|_X^2$ converges with the same rate as $1 - LS(f;\mathfrak{u}_h)/\|\mathfrak{u} - \mathfrak{u}_h\|_X^2$.

Experiment 2 (Asymptotic exactness with unknown solution). This experiment investigates the properties of the discrete solution $\mathfrak{u}_h = (u_h, p_h) = \arg\min_{x_h \in X_h} LS(f; x_h)$ to the Poisson model problem with right-hand side $f \equiv 1$ on the L-shaped domain $\Omega = (-1,1)^2 \setminus [0,1)^2$ and $X_h = S_0^k(\mathcal{T}) \times RT_{k-1}(\mathcal{T})$ for $k = 1, 2, 3$. A reference solution $\mathfrak{u}_{\mathrm{ref}} = \arg\min_{x_{\mathrm{ref}} \in X_{\mathrm{ref}}} LS(f; x_{\mathrm{ref}})$ with $X_h \subset X_{\mathrm{ref}} \subset X$ approximates the unknown exact solution $\mathfrak{u} \in X$. Algorithm 2 displays the computation of $\mathfrak{u}_{\mathrm{ref}}$ with an adaptive algorithm that produces a sequence of approximations $(\mathfrak{u}_{\mathrm{ref}}^\ell)_{\ell \in \mathbb{N}}$. The algorithm stops, if the refinement of the mesh does not change the solution significantly, that is

$$\left| 1 - \frac{|1 - LS(f;\mathfrak{u}_h)/\|\mathfrak{u}_{\mathrm{ref}}^{\ell-1} - \mathfrak{u}_h\|_X^2|}{|1 - LS(f;\mathfrak{u}_h)/\|\mathfrak{u}_{\mathrm{ref}}^{\ell} - \mathfrak{u}_h\|_X^2|} \right| \leq 0.01. \tag{3.44}$$

Figure 3.2 displays the result of the experiment. It indicates $0 < 1 - LS(f;\mathfrak{u}_h)/\|\mathfrak{u}_{\mathrm{ref}} - \mathfrak{u}_h\|_X^2 = \mathcal{O}(h_{\max}^{4/3}) = \mathcal{O}(\text{ndof}^{-2/3})$ for uniform mesh refinements. This suggests that the rate of convergence equals the squared elliptic regularity of the domain (which reads $2/3$ for the L-shaped domain Ω [Dau88, Chap. 6]). The right-hand side in Figure 3.2 plots the results of the experiment with the adaptive refinement strategy from Algorithm 3 on page 137 with bulk parameter $\Theta = 0.3$ and refinement indicator $\eta^2(T) := \|\nabla u_h - p_h\|_{L^2(T)}^2 +$

Algorithm 2: Computation of $\mathfrak{u}_{\mathrm{ref}}$

Input: $k = 1, 2, 3$, $\mathfrak{u}_h =: \mathfrak{u}_h^0 \in S_0^k(\mathcal{T}) \times RT_{k-1}(\mathcal{T})$, $\mathcal{T}_1 := \mathcal{T}$, $f \equiv 1$

for $\ell = 1, 2, \ldots$ **do**
- set $X_{\mathrm{ref}}^\ell := S_0^{k+3}(\mathcal{T}_\ell) \times RT_{k+2}(\mathcal{T}_\ell)$ and $(u_h, p_h) = \mathfrak{u}_{\mathrm{ref}}^\ell = \arg\min_{x_h \in X_{\mathrm{ref}}^\ell} LS(f; x_h)$;
- **if** *stopping criterion (3.44)* **then**
  - **break**;
- **else**
  - compute an adaptively refined mesh $\mathcal{T}_{\ell+1}$ with Algorithm 3 on page 137 with $\Theta = 0.3$ and $\eta^2(T) := \|\nabla u_h - p_h\|_{L^2(T)}^2 + \|\operatorname{div} p_h + f\|_{L^2(T)}^2$ for all $T \in \mathcal{T}_\ell$;
- **end**

end

Output: the reference solution $\mathfrak{u}_{\mathrm{ref}} := \mathfrak{u}_{\mathrm{ref}}^\ell \in X_{\mathrm{ref}}^\ell =: X_{\mathrm{ref}}$

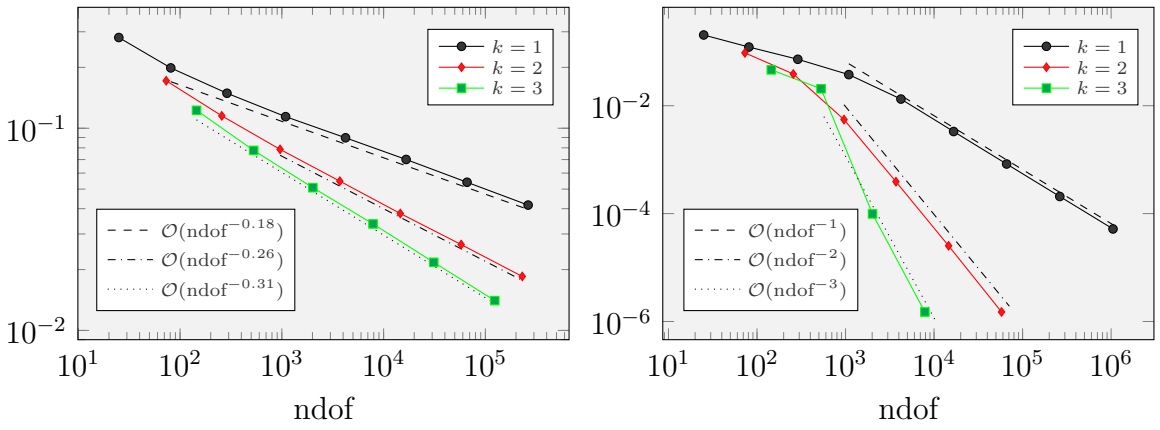


Figure 3.3: Distance $C(X_h) - 1$ with $N = 600$ (left) and $C(X_h) - \mu_{12}^{-1}$ with $N = 11$ in Experiment 3 with polynomial degrees $k = 1, 2, 3$

$\|\operatorname{div} p_h + f\|_{L^2(T)}^2$ for all $T \in \mathcal{T}$. Since the accuracy of the solutions $\mathfrak{u}_h$ to the LSFEM with polynomial degree $k = 2, 3$ do not allow for the computation of sufficiently accurate approximations $\mathfrak{u}_{\mathrm{ref}}$ with (3.44), the convergence history plot on the right-hand side of Figure 3.2 plots only the results for $k = 1$. The adaptive algorithm results in a strong refinement of the re-entrant corner and $\|\mathfrak{u} - \mathfrak{u}_h\|_a^2 = \mathcal{O}(\mathrm{ndof}^{-1}) = 1 - LS(f; \mathfrak{u}_h)/\|\mathfrak{u}_{\mathrm{ref}} - \mathfrak{u}_h\|_X^2$ for $k = 1$. The ratio $LS(f; \mathfrak{u}_h)/\|\mathfrak{u}_{\mathrm{ref}} - \mathfrak{u}_h\|_X^2$ is smaller than one in all computations.

Experiment 3 (Improved GUBs on square domain). This experiment exploits the improvement of the reliability constant $C(X_h)$ with the three-stage algorithm from Section 3.1.2 for the Poisson model problem on the unit square domain $\Omega = (0, 1)^2$. The discrete space reads $X_h := S_0^k(\mathcal{T}) \times RT_{k-1}(\mathcal{T})$ with $k = 1, 2, 3$ and uniformly refined triangulations $\mathcal{T}$ of Ω. The eigenvalues $0 < \mu_1 = \mu_1^{\mathrm{low}} \leq \cdots \leq \mu_{N+1} = \mu_{N+1}^{\mathrm{low}}$ from Stage 1 result from the identity $\mu_j = 1 - (1 + \lambda_j)^{-1/2}$ with the known Dirichlet eigenvalues $\lambda_1 = 2\pi^2, \lambda_2 = 5\pi^2, \ldots$ of the Laplace operator. Appendix A.1.2 explains the computation of the upper eigenvalue bounds $\mu_{h,1}^{\mathrm{up}} \leq \cdots \leq \mu_{h,N}^{\mathrm{up}}$ from Stage 2. Figure 3.3 displays the distance $C(X_h) - 1$ for $N = 600$. Since $n \ll N$ results in the minimum value $C(X_h)$ in

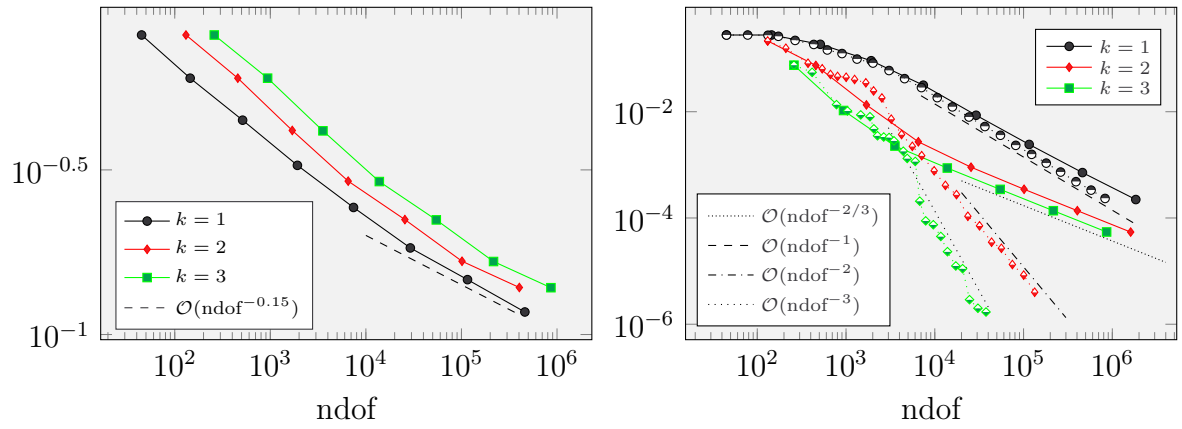


Figure 3.4: Distance $C(X_h) - 1$ with a) and $N = 100$ (left) and $C(X_h) - \mu_8^{-1}$ with b) and $N = 7$ (right) in Experiment 4

Algorithm 1 on page 23, the result should be similar to $N = \dim X_h$. In other words, the experiment leads to the best possible $C(X_h)$ with the three-stage algorithm. The distance $C(X_h) - 1$ seems to converge with very small rates. The rates increase as the polynomial degrees increase.

The right-hand side of Figure 3.3 displays the convergence of $C(X_h) - \mu_{12}$ with reliability constant $C(X_h)$ from the three-stage algorithm with $N = 11$. It shows the predicted rates of convergence $C(X_h) - 1 = \mathcal{O}(h_{\max}^{2k}) = \mathcal{O}(\text{ndof}^{-k})$ from Remark 3.1.12 with some small pre-asymptotic regime for all $k = 1, 2, 3$.

Experiment 4 (Improved GUBs on L-shaped domain). This experiment exploits the improvement of the reliability constant $C(X_h)$ with the three-stage algorithm from Section 3.1.2 for the Poisson model problem on the L-shaped domain $\Omega = (-1, 1)^2 \setminus [0, 1)^2$. Since the exact Dirichlet eigenvalues $\lambda_1, \lambda_2, \dots$ of the Laplace operator are unknown, the computation of lower eigenvalue bounds for $\mu_j = 1 - (1 + \lambda_j)^{1/2}$ utilizes either

a) the algorithm from [CG14b] (see Appendix A.3.2 for details), which applies on the fly on each triangulation $\mathcal{T}$ and computes lower eigenvalue bounds $\lambda_j^{\text{low}} \leq \lambda_j$ for all $j = 1, \dots, N$ and $N \leq \dim S_0^1(\mathcal{T})$, or

b) accurate approximations of $\lambda_1, \dots, \lambda_8$ from [TB06, Sec. 2] and [Gal14a, Chap. 9.2].

Figure 3.3 displays the result for $X_h = S_0^k(\mathcal{T}) \times RT_{k-1}(\mathcal{T})$ with $k = 1, 2, 3$ and uniformly (left and right) and adaptively (right) refined triangulations $\mathcal{T}$ of Ω with the refinement strategy from Experiment 2 for $\mathfrak{u}_h = (u_h, p_h) = \arg\min_{x_h \in X_h} LS(f; x_h)$ with $f \equiv 1$. The left-hand side of Figure 3.3 shows the convergence history plot of the distance $C(X_h) - 1$, where $C(X_h)$ utilizes the lower eigenvalue bounds from a) and $N = 100$. It indicates the poor convergence speed $C(X_h) - 1 = \mathcal{O}(\text{ndof}^{-0.15})$ for all $k = 1, 2, 3$. Since the finer underlying triangulations $\mathcal{T}$ for $k = 1$ allow for better approximations of the lower eigenvalue bounds with the Crouzeix-Raviart-FEM, the distance $C(X_h) - 1$ is smaller for the polynomial degree $k = 1$ than for $k = 2, 3$.

The right-hand side of Figure 3.3 shows the convergence history plot of $C(X_h) - 1$, where $C(X_h)$ utilizes the eigenvalues $\mu_1, \dots, \mu_8$ from b) and $N = 7$. It indicates the expected speed of convergence $C(X_h) - \mu_8^{-1} = \mathcal{O}(\text{ndof}^{-2/3})$ from Remark 3.1.12 for uniform mesh refinements (solid lines) and $k = 1, 2, 3$. Adaptive mesh refinements (dotted lines) improve

	$\mathbb{M}$	A	γ	D	D^*	V	Σ
Poisson	$\mathbb{R}^d$	id	0	∇	$-\mathrm{div}$	$H^1_0(\Omega)$	$H(\mathrm{div},\Omega)$
Helmholtz	$\mathbb{R}^d$	id	ω^2	∇	$-\mathrm{div}$	$H^1_0(\Omega)$	$H(\mathrm{div},\Omega)$
Elasticity	$\mathbb{R}^{d\times d}$	$\mathbb{C}$	0	$\varepsilon(\bullet)$	$-\mathrm{div}$	$H^1_0(\Omega;\mathbb{R}^d)$	$H(\mathrm{div},\Omega;\mathbb{R}^{d\times d})$
Maxwell	$\mathbb{R}^3$	id	ω^2	curl	curl	$H_0(\mathrm{curl},\Omega)$	$H(\mathrm{curl},\Omega)$

Table 3.1: Notation in Section 3.2.2

the rate of convergence significantly. Probably, the strong refinement of the re-entrant allows for better approximations of the eigenfunctions $\psi_1,\dots,\psi_7\in X$ which correspond to the eigenvalues $\mu_1,\dots,\mu_7$ and so the eigenvalue error in Remark 3.1.12 reads $\delta=\mathcal{O}(\mathrm{ndof}^{-k})$ for $k=1,2,3$.

Discussion. The asymptotic exactness result in Theorem 3.1.7 states the convergence $LS(f;\mathfrak{u}_h)/\|\mathfrak{u}-\mathfrak{u}_h\|^2_X$ without a convergence rate. Experiment 1–2 suggest that this rate is (almost) independent of the polynomial degree of the discrete ansatz space X_h for uniformly refined meshes, but depends on the elliptic regularity of the domain, that is, $LS(f;\mathfrak{u}_h)/\|\mathfrak{u}-\mathfrak{u}_h\|^2_X=1+\mathcal{O}(h^{2s}_{\max})$ where $s\in(0,1]$ is the elliptic regularity of the domain. Adaptive mesh refinements can accelerate the convergence of the ratio $LS(f;\mathfrak{u}_h)/\|\mathfrak{u}-\mathfrak{u}_h\|^2_X$ towards one (see the right-hand side of Figure 3.2).

Remark 3.1.12 with $N\to\infty$, exact arithmetic, and convergent (or exact) lower eigenvalue bounds implies the convergence of the reliability constant $C(X_h)$ from the three-stage algorithm towards one. Experiment 3 indicates a poor rate of convergence. This poor convergence and the enormous computational effort suggest that practical applications cannot aim for a reliability constant $C(X_h)$ arbitrarily close to one. Instead, practical applications should compute reliability constants $C(X_h)$ with fixed and small number $N\in\mathbb{N}$. These computations result in reliability constants $C(X_h)$ which converge towards μ^{-1}_{N+1} (see Remark 3.1.12 and Experiment 3–4) with moderate computational effort (the time for computing $C(X_h)$ with $N=11$ and $N=7$ in Experiment 3 and 4 is 3–6 times larger than the time for computing the discrete solution $\mathfrak{u}_h=\arg\min_{x_h\in X_h} LS(f;x_h)$) and good rates of convergence. Unfortunately, the improvement of the reliability constant $C(X_h)$ compared to α^{-1} is very small: $\alpha^{-1}=1.281$ and $\mu^{-1}_{12}=1.076$ in Experiment 3 and $\alpha^{-1}=1.442$ and $\mu^{-1}_8=1.164$ in Experiment 4. Thus, the improved GUB $C(X_h)LS(f;\mathfrak{u}_h)$ does not lead to significant improvement of the natural GUB $\alpha^{-1}LS(f;\mathfrak{u}_h)$ for the Poisson model problem. However, the next section underlines the practicability of the improved GUB.

3.2.2 Elasticity, Helmholtz, Maxwell

This section generalizes the techniques from Section 3.2.1 to validate (**H1**)–(**H4**) for the LSFEMs for the Helmholtz equation from [CLMM94], the linear elasticity from [CKS05], and the time-harmonic Maxwell equations from [BG09, Sec. 6.4]. Table 3.1 introduces abstract operators which compile these problems. The table contains the positive definite isomorphism $A=A^{1/2}\circ A^{1/2}$, which maps the subspace $\mathbb{M}\subseteq\mathbb{R}^{m\times n}$ with $m,n\in\mathbb{N}$ onto $\mathbb{M}$. Furthermore, the table contains the linear differential operator D, which maps the real Hilbert space V with norm $\|\bullet\|^2_V=\|\bullet\|^2_{L^2(\Omega)}+\|A^{1/2}D\bullet\|^2_{L^2(\Omega)}$ onto a closed subset of $L^2(\Omega;\mathbb{M})$. Since $D:V\to L^2(\Omega;\mathbb{M})$ is bounded, its kernel $\ker D$ is closed and so there exists

an orthogonal complement $W \subset V$ with $W \perp_V \ker D$ and $V = W \oplus \ker D$. Theorem 2.2.1, 2.3.4, and 2.4.2 show that there exist countably many eigenpairs $(\lambda_j, \phi_j) \in \mathbb{R} \times W \setminus \{0\}$ with $D^*AD\phi_j = \lambda_j \phi_j$ for $j \in \mathbb{N}$, that is

$$(AD\phi_j, Dv)_{L^2(\Omega)} = \lambda_j (\phi_j, v)_{L^2(\Omega)} \qquad \text{for all } v \in V. \tag{3.45}$$

The eigenfunctions $(\phi_j)_{j\in\mathbb{N}}$ form a basis of $W = \overline{\operatorname{span}\{\phi_j \mid j \in \mathbb{N}\}}^{\|\bullet\|_V}$ and are orthonormal in the sense that $(\phi_j, \phi_k)_{L^2(\Omega)} = \delta_{jk}$ and $(AD\phi_j, D\phi_k)_{L^2(\Omega)} = \lambda_j \delta_{jk}$ for all $j, k \in \mathbb{N}$. Moreover, $0 < \lambda_1 \le \lambda_2 \le \dots$ with $\lim_{j\to\infty} \lambda_j = \infty$. The spaces $\Sigma := \{\tau \in L^2(\Omega; \mathbb{M}) \mid D^*\tau \in L^2(\Omega; \mathbb{R}^m)\}$ from Table 3.1 with squared norm $\|\bullet\|_\Sigma^2 := \|A^{-1/2}\bullet\|_{L^2(\Omega)}^2 + \|D^*\bullet\|_{L^2(\Omega)}^2$ are Hilbert spaces. Since $D^* : \Sigma \to L^2(\Omega; \mathbb{R}^m)$ is linear and bounded, the kernel $\ker D^*$ is a closed subspace of Σ. Given $f \in L^2(\Omega; \mathbb{R}^m)$, the problems seek the solution $\mathfrak{u} = (u, \sigma) \in X := V \times \Sigma$ to

$$A^{-1/2}\sigma - A^{1/2}Du = 0 \qquad \text{and} \qquad D^*\sigma - \gamma u = f. \tag{3.46}$$

Suppose that problem (3.46) is well posed, that is either the kernel $\ker D = \{0\}$ is trivial and $\gamma = 0$ or $\gamma \in (0, \infty) \setminus \{\lambda_1, \lambda_2, \dots\}$ (see Theorem 2.1.3, 2.2.2, 2.3.3, and 2.4.3). Then the solution $\mathfrak{u} = (u, \sigma) \in X$ is the unique minimizer of the least-squares functional

$$LS(f; v, \tau) := \|A^{-1/2}\tau - A^{1/2}Dv\|_{L^2(\Omega)}^2 + \|D^*\tau - \gamma v - f\|_{L^2(\Omega)}^2 \qquad \text{over all } (v, \tau) \in X.$$

Moreover, (3.46) shows $LS(f; v, \tau) = \|(u, \sigma) - (v, \tau)\|_a^2$ with squared norm

$$\|(v, \tau)\|_a^2 := \|A^{-1/2}\tau - A^{1/2}Dv\|_{L^2(\Omega)}^2 + \|D^*\tau - \gamma v\|_{L^2(\Omega)}^2 \qquad \text{for all } (v, \tau) \in X. \tag{3.47}$$

Let the inner product $a(\bullet, \bullet)$ induce the norm $\|\bullet\|_a$ and let the inner product $(\bullet, \bullet)_X$ induce the norm $\|\bullet\|_X$ with, for all $(v, \tau) \in X$,

$$\|(v, \tau)\|_X^2 := \gamma^2 \|A^{1/2}v\|_{L^2(\Omega)}^2 + \|Dv\|_{L^2(\Omega)}^2 + \|A^{-1/2}\tau\|_{L^2(\Omega)}^2 + \|D^*\tau\|_{L^2(\Omega)}^2. \tag{3.48}$$

For all $j \in \mathbb{N}$ define $\nu_j := \lambda_j(\gamma + 1)^2/((\lambda_j + 1)(\gamma^2 + \lambda_j))$ and set

$$\mu_0 := 1 \quad \text{and} \quad \psi_0 \in \ker D \times \ker D^* \subset X, \tag{3.49a}$$

$$\mu_{2j-1} := 1 - \nu_j^{1/2} \quad \text{and} \quad \psi_{2j-1} := \Big((\lambda_j^2 + \lambda_j)^{1/2}(\gamma^2 + \lambda_j)^{-1/2}\phi_j, AD\phi_j\Big) \in X, \tag{3.49b}$$

$$\mu_{2j} := 1 + \nu_j^{1/2} \quad \text{and} \quad \psi_{2j} := \Big((\lambda_j^2 + \lambda_j)^{1/2}(\gamma^2 + \lambda_j)^{-1/2}\phi_j, -AD\phi_j\Big) \in X. \tag{3.49c}$$

Theorem 3.2.4 (Eigenvalues of $a(\bullet, \bullet)$ and $(\bullet, \bullet)_X$)**.** *The formulae in (3.49) define eigenpairs $(\mu_j, \psi_j) \in \mathbb{R} \times X$ with*

$$a(\psi_j, x) = \mu_j (\psi_j, x)_X \qquad \textit{for all } j \in \mathbb{N}_0 \textit{ and } x \in X. \tag{3.50}$$

The proof of (3.50) utilizes the following Helmholtz decomposition.

Lemma 3.2.5 (Helmholtz decomposition in Σ)**.** *The splits*

$$L^2(\Omega; \mathbb{M}) = AD(V) \oplus \ker D^* \qquad \textit{and} \qquad \Sigma = \Big(AD(V) \cap \Sigma\Big) \oplus \ker D^* \tag{3.51}$$

are orthogonal with respect to the inner product $(A^{-1}\bullet, \bullet)_{L^2(\Omega)}$.

Proof. Step 1 (Decomposition of $L^2(\Omega;\mathbb{M})$). Since the norm $\|A^{1/2}D\bullet\|_{L^2(\Omega)}$ in $W \subset V = W \oplus \ker D$ is equivalent to $\|\bullet\|_V$, $\big(W, (AD\bullet, D\bullet)_{L^2(\Omega)}\big)$ is a Hilbert space. Given a function $\tau \in L^2(\Omega;\mathbb{M})$, the unique Riesz representation $\xi \in W$ satisfies

$$(AD\xi, Dw)_{L^2(\Omega)} = (\tau, Dw)_{L^2(\Omega)} \qquad \text{for all } w \in W.$$

Define $\tau_0 := \tau - AD\xi$ with $(\tau_0, Dw)_{L^2(\Omega)} = (\tau, Dw)_{L^2(\Omega)} - (AD\xi, Dw)_{L^2(\Omega)} = 0$ for all $w \in W$, whence $\tau_0 \in \ker D^*$. Since $\tau_0 \in \ker D^*$ and $(A^{-1}ADv, \tau_0)_{L^2(\Omega)} = (v, D^*\tau_0)_{L^2(\Omega)} = 0$ for all $v \in V$, the split is orthogonal.

Step 2 (Decomposition of Σ). Given $\tau \in \Sigma \subset L^2(\Omega;\mathbb{M})$, Step 1 leads to functions $\xi \in V$ and $\tau_0 \in \ker D^*$ with $\tau = AD\xi + \tau_0$ and $(A^{-1}AD\xi, \tau_0)_{L^2(\Omega)} = 0$. Since $\tau, \tau_0 \in \Sigma$ and Σ is a vector space, $AD\xi = \tau - \tau_0 \in \Sigma$. □

Proof of Theorem 3.2.4. Step 1 (Decomposition of the bilinear forms). Given functions $(v,\tau), (w,\chi) \in X$, (3.51) leads to functions $\vartheta, \xi \in W$ and $\tau_0, \chi_0 \in \ker D^*$ with $AD\vartheta, AD\xi \in \Sigma$, $\tau = AD\vartheta + \tau_0$, and $\chi = AD\xi + \chi_0$. Since $\operatorname{span}\{\phi_j \mid j \in \mathbb{N}\}$ is dense in W with respect to the norm $\|\bullet\|_V$ and $V = W \oplus \ker D$, there exist coefficients $v_j, w_j, \vartheta_j, \xi_j \in \mathbb{R}$ for $j \in \mathbb{N}$ and functions $v_0, w_0 \in \ker D$ with

$$v = v_0 + \sum_{j\in\mathbb{N}} v_j\phi_j, \quad w = w_0 + \sum_{j\in\mathbb{N}} w_j\phi_j, \quad \vartheta = \sum_{j\in\mathbb{N}} \vartheta_j\phi_j, \quad \text{and} \quad \xi = \sum_{j\in\mathbb{N}} \xi_j\phi_j.$$

The orthogonality of the eigenfunctions results in the formal calculation

$$\begin{aligned}
a(v,\tau;w,\chi) &= (A^{1/2}D(v-\vartheta) - A^{-1/2}\tau_0, A^{1/2}D(w-\xi) - A^{-1/2}\chi_0)_{L^2(\Omega)} \\
&\quad + (\gamma v - D^*AD\vartheta, \gamma w - D^*AD\xi)_{L^2(\Omega)} \\
&= (\sum_{j\in\mathbb{N}} (v_j - \vartheta_j)A^{1/2}D\phi_j, \sum_{k\in\mathbb{N}} (w_k - \xi_k)A^{1/2}D\phi_k)_{L^2(\Omega)} \\
&\quad + (\sum_{j\in\mathbb{N}} (\gamma v_j - \lambda_j\vartheta_j)\phi_j, \sum_{k\in\mathbb{N}} (\gamma w_k - \lambda_k\xi_k)\phi_k)_{L^2(\Omega)} \\
&\quad + (A^{-1/2}\tau_0, A^{-1/2}\chi_0)_{L^2(\Omega)} + \gamma^2(v_0, w_0)_{L^2(\Omega)} \\
&= \sum_{j\in\mathbb{N}} \begin{pmatrix} v_j \\ \vartheta_j \end{pmatrix} \cdot \begin{pmatrix} \lambda_j + \gamma^2 & -\lambda_j - \gamma\lambda_j \\ -\lambda_j - \gamma\lambda_j & \lambda_j + \lambda_j^2 \end{pmatrix} \begin{pmatrix} w_j \\ \xi_j \end{pmatrix} \\
&\quad + (A^{-1/2}\tau_0, A^{-1/2}\chi_0)_{L^2(\Omega)} + \gamma^2(v_0, w_0)_{L^2(\Omega)}.
\end{aligned}$$

Similar arguments lead to

$$\begin{aligned}
(v,\tau;w,\chi)_X = &\sum_{j\in\mathbb{N}} \begin{pmatrix} v_j \\ \vartheta_j \end{pmatrix} \cdot \begin{pmatrix} \lambda_j + \gamma^2 & 0 \\ 0 & \lambda_j + \lambda_j^2 \end{pmatrix} \begin{pmatrix} w_j \\ \xi_j \end{pmatrix} \\
&+ (A^{-1/2}\tau_0, A^{-1/2}\chi_0)_{L^2(\Omega)} + \gamma^2(v_0, w_0)_{L^2(\Omega)}.
\end{aligned}$$

Step 2 (Computation of eigenpairs). The decomposition of $a(\bullet,\bullet)$ and $(\bullet,\bullet)_X$ in Step 1 shows that $\mu_0 = 1$ satisfies (3.50) for all elements ψ_0 in $\ker D \times \ker D^*$. Moreover, the

Problem	LS Eigenvalues
Poisson Elasticity	$\mu_{2j-1} = 1 - (\lambda_j + 1)^{-1/2}$ $\mu_{2j} = 1 + (\lambda_j + 1)^{-1/2}$
Helmholtz Maxwell	$\mu_{2j-1} = 1 - (\lambda_j(\omega^2+1)^2(\lambda_j+1)^{-1}(\omega^4+\lambda_j)^{-1})^{1/2}$ $\mu_{2j} = 1 + (\lambda_j(\omega^2+1)^2(\lambda_j+1)^{-1}(\omega^4+\lambda_j)^{-1})^{1/2}$

Table 3.2: Eigenvalues μ_j with (3.50) in dependence of λ_j from (3.45) for all $j \in \mathbb{N}$

decomposition leads for all $j \in \mathbb{N}$ and all $(w, \chi) \in X$ with decomposition as in Step 1 to

$$\begin{aligned} a(\psi_{2j-1}; w, \chi) &= \begin{pmatrix} (\lambda_j^2 + \lambda_j)^{1/2}(\gamma^2 + \lambda_j)^{-1/2} \\ 1 \end{pmatrix} \cdot \begin{pmatrix} \lambda_j + \gamma^2 & -\lambda_j - \gamma\lambda_j \\ -\lambda_j - \gamma\lambda_j & \lambda_j + \lambda_j^2 \end{pmatrix} \begin{pmatrix} w_j \\ \xi_j \end{pmatrix} \\ &= \mu_{2j-1} \begin{pmatrix} (\lambda_j^2 + \lambda_j)^{1/2}(\gamma^2 + \lambda_j)^{-1/2} \\ 1 \end{pmatrix} \cdot \begin{pmatrix} \lambda_j + \gamma^2 & 0 \\ 0 & \lambda_j + \lambda_j^2 \end{pmatrix} \begin{pmatrix} w_j \\ \xi_j \end{pmatrix} \\ &= \mu_{2j-1} \, (\psi_{2j-1}; w, \chi)_X. \end{aligned}$$

Analogously, $a(\psi_{2j}; w, \chi) = \mu_{2j} \, (\psi_{2j}; w, \chi)_X$ follows for all $j \in \mathbb{N}$ and $(w, \chi) \in X$. □

Theorem 3.2.6 (Properties (**H1**)–(**H4**))**.** *The inner products $a(\bullet, \bullet)$ and $(\bullet, \bullet)_X$ in $X = V \times \Sigma$, which induce the norms $\|\bullet\|_a$ and $\|\bullet\|_X$ from (3.47)–(3.48), satisfy* (**H1**)–(**H4**)*.*

Proof. Step 1 (Proof of (**H1**)*).* The eigenvalues $\mu_0, \mu_1, \ldots$ from (3.49) lead to countably many pairwise distinct numbers $\mu(0) = 1, \mu(1), \mu(2), \ldots$ with $\{\mu_k \mid k \in \mathbb{N}_0\} = \{\mu(j) \mid j \in \mathbb{N}_0\}$. Theorem 3.2.4 proves that the closed subspaces $E(\mu(0)) := \ker D \times \ker D^*$ and $E(\mu(j)) := \operatorname{span}\{\psi_k \mid k \in \mathbb{N} \text{ and } \mu_k = \mu(j)\}$ for all $j \in \mathbb{N}$ with eigenfunction $\psi_k \in X$ from (3.49) satisfy (**H1**).

Step 2 (Proof of (**H4**)*).* A simple calculation shows that μ_j from (3.49) and so $\mu(j)$ tend to one as j (and so λ_j) tends to infinity.

Step 3 (Proof of (**H2**)*).* The eigenspace $E(\mu(j))$ is the linear hull of all ψ_k with $\mu_k = \mu(j)$. The eigenfunctions $\psi_1, \psi_2, \ldots$ from (3.49) are linearly independent and so $\dim E(\mu(j)) = |\{k \in \mathbb{N} \mid \mu_k = \mu(j)\}|$ with the counting measure $|\bullet|$. Since $\lim_{k\to\infty} \mu_k = 1$ and $\mu(j) \neq 1$, the number $|\{k \in \mathbb{N} \mid \mu_k = \mu(j)\}|$ is finite for all $j \in \mathbb{N}$.

Step 4 (Proof of (**H3**)*).* The model problems satisfy either $\gamma \neq 0$ or $\gamma = 0$ and $\ker D = \{0\}$. If $\gamma \neq 0$, it holds

$$\min\{1, \gamma^2\}(\|v\|_V^2 + \|\tau\|_\Sigma^2) \leq \|(v, \tau)\|_X^2 \leq \max\{1, \gamma^2\}(\|v\|_V^2 + \|\tau\|_\Sigma^2) \quad \text{for all } (v, \tau) \in \Sigma.$$

If $\ker D = \{0\}$ and $\gamma = 0$, it holds

$$(1 + \lambda_1^{-1})^{-1}(\|v\|_V^2 + \|\tau\|_\Sigma^2) \leq \|(v, \tau)\|_X^2 \leq \|v\|_V^2 + \|\tau\|_\Sigma^2 \quad \text{for all } (v, \tau) \in \Sigma.$$

Hence, the norms $\|\bullet\|_X$ and $\|\bullet\|_{V\times\Sigma} := (\|\bullet\|_V^2 + \|\bullet\|_\Sigma^2)^{1/2}$ are equivalent. Since $\operatorname{span}\{\psi_j \mid j \in \mathbb{N}\}$ is dense in W with respect to the norm $\|\bullet\|_V$ in V, the element $(\phi_j, 0) \in$

$\operatorname{span}\{E(\mu_{2j-1}), E(\mu_{2j})\}$ for all $j \in \mathbb{N}$, the kernel $\ker D \times \{0\} \subset E(\mu(0))$, and $V = W \oplus \ker D$, the equivalence of the norms $\|\bullet\|_X$ and $\|\bullet\|_{V\times\Sigma}$ implies

$$V \times \{0\} = \ker D \times \{0\} \oplus \overline{\operatorname{span}\{(\phi_j, 0) \mid j \in \mathbb{N}\}}^{\|\bullet\|_X} \subset \overline{\operatorname{span}\{E(\mu(j)) \mid j \in \mathbb{N}_0\}}^{\|\bullet\|_X}.$$

The density of $\operatorname{span}\{AD\psi_1, AD\psi_2, \dots\}$ in $AD(V) \cap \Sigma$ with respect to the norm $\|\bullet\|_\Sigma$ implies $(0, AD\phi_j) \in \operatorname{span}\{E(\mu_{2j-1}), E(\mu_{2j})\}$ for all $j \in \mathbb{N}$. This, $\{0\} \times \ker D^* \subset E(\mu(0))$, and the equivalence of the norms $\|\bullet\|_X$ and $\|\bullet\|_{V\times\Sigma}$ show

$$\{0\} \times \Sigma = \{0\} \times \ker D^* \oplus \overline{\operatorname{span}\{(0, AD\phi_j) \mid j \in \mathbb{N}\}}^{\|\bullet\|_X} \subset \overline{\operatorname{span}\{E(\mu(j)) \mid j \in \mathbb{N}_0\}}^{\|\bullet\|_X}. \square$$

Remark 3.2.7. *Table 3.2 shows that small eigenvalues λ of the differential operator D^*AD cause small and large eigenvalues μ in the Poisson model problem and the linear elasticity. However, small and large eigenvalues μ for the Maxwell and Helmholtz equations result not only from the size of the eigenvalues λ but also from the distance $|\lambda - \omega^2|$ to the frequency ω. Figure 3.5 presents a corresponding example with a huge pre-asymptotic regime caused by the necessity to resolve the eigenfunctions ϕ_j with eigenfrequency λ_j close to ω^2 sufficiently well.*

Remark 3.2.8 (Generalizations)**.** *The detailed analysis behind Table 3.2 is performed for the four model problems but can be extended to other norms and problems. For example, Section 3.2.3 extends the analysis to the Stokes problem and the supplementary material of [CS18] extends the analysis to the LSFEM for the Helmholtz equations with the alternative inner product, for all $(v, \tau), (w, \chi) \in H^1_0(\Omega) \times H(\operatorname{div}, \Omega)$,*

$$(v, \tau; w, \chi)_{X_0} := (\nabla v, \nabla w)_{L^2(\Omega)} + (\tau, \chi)_{L^2(\Omega)} + (\operatorname{div} \tau, \operatorname{div} \chi)_{L^2(\Omega)}.$$

Numerical experiments

Theorem 3.2.6 proves (**H1**)–(**H4**) and so the three-stage algorithm in Section 3.1.2 results in an improved reliability constant $C(X_h)$ for the model problems of this section. The numerical experiments of this section investigate the practicability of the improved GUB $C(X_h)LS(f; \mathfrak{u}_h)$ for the Helmholtz and Maxwell equations. A detailed description of the implemented routines in FEniCS is postponed to Appendix A.1.

Experiment 1 (Improved GUBs for Helmholtz). This experiment investigates the three-stage algorithm on page 23 for the Helmholtz equation on the square domain $\Omega = (0, 1)^2$ with $N = 11, 60$ and discrete space $X_h = S^k_0(\mathcal{T}) \times RT_{k-1}(\mathcal{T})$ from (3.40) with $k = 1, 2, 3$. It utilizes the known Dirichlet eigenvalues $\lambda_1 = 2\pi^2, \lambda_2 = 5\pi^2, \dots$ of the Laplace operator to compute exact eigenvalues $\mu^{\text{low}}_1 = \mu_1 \le \mu^{\text{low}}_2 = \mu_2 \le \dots$ with the formula in Table 3.2. The upper eigenvalue bounds $\mu_{h,1} \le \mu^{\text{up}}_{h,1} \le \mu_{h,2} \le \dots \le \mu_{h,N} \le \mu^{\text{up}}_{h,N}$ in Stage 2 of the three-stage algorithm result from the algorithm in Appendix A.1.2. Table 3.3 displays the resulting reliability constants $\alpha^{-1} = \mu_1^{-1}$ and $C(X_h)$ for the frequencies $\omega = 0, 1, \dots, 10$. The improvement of the reliability constant is often significant, especially for moderate frequencies and large polynomial degrees. However, the three-stage algorithm does not improve the reliability constant for $k = 1$ and $\omega = 7, 10$. Experiment 2 investigates this phenomenon.

		$C(X_h), k=1$		$C(X_h), k=2$		$C(X_h), k=3$	
ω	α^{-1}	$N=11$	$N=73$	$N=11$	$N=73$	$N=11$	$N=73$
0	1.28137	1.07663	1.04168	1.07642	1.03088	1.07642	1.03143
1	1.74963	1.16552	1.08717	1.16500	1.06906	1.16500	1.06902
2	5.43345	1.52114	1.25047	1.51840	1.18129	1.51840	1.20061
3	35.7235	2.51502	1.65317	2.48648	1.45001	2.48648	1.44061
4	817.514	11.4324	9.02788	4.91147	1.95401	4.91085	1.90290
5	965.782	20.2394	14.3835	10.4558	2.77875	10.4547	2.77388
6	759.849	41.5941	29.2401	22.2419	4.43888	22.2375	4.42467
7	2037167	1992453	1992453	18592.6	18553.7	76.4978	37.0047
8	2983.90	1023.41	1007.76	95.5873	11.0956	95.2386	10.6419
9	254356	232344	232344	999.940	821.566	199.882	19.0285
10	1184260	1164859	1164859	23055.6	22853.1	465.847	100.023

Table 3.3: Reliability constants α^{-1} and $C(X_h)$ with the three-stage algorithm for the Helmholtz equation with uniformly refined meshes and ndof = 263169 for $k=1$, ndof = 230401 for $k=2$, and ndof = 123649 for $k=3$ in Experiment 1

Experiment 2 (Difficulties of LSFEM). Table 3.3 displays large reliability constants $C(X_h) \leq \alpha^{-1}$ for squared frequencies ω^2 close to Dirichlet eigenvalues of the Laplace operator. The large reliability constants $C(X_h)$ result from the inaccurate approximation of the smallest eigenvalue μ_1 by the smallest discrete eigenvalue $\mu_{h,1}$. More precisely, the formula in (3.22) with $m = 1, \ldots, \dim X_h$ implies

$$\frac{\mu_{h,1}-\mu_1}{\mu_{h,1}\mu_1} \leq \mu_{m+1}^{-1} + \frac{\mu_{h,1}-\mu_1}{\mu_{h,1}\mu_1}\frac{\mu_{m+1}-\mu_1}{\mu_2-\mu_1}\frac{\mu_2}{\mu_{m+1}} \leq C(X_h). \tag{3.52}$$

Thus, a small reliability constant $C(X_h)$ requires a small error $(\mu_{h,1}-\mu_1)\mu_{h,1}^{-1}\mu_1^{-1}$. However, the denominator $\mu_{h,1}\mu_1 \approx \mu_1^2$ is tiny for frequencies ω^2 close to Dirichlet eigenvalues of the Laplace operator and so a small reliability constant requires a very accurate approximation $\mu_{h,1}$ of μ_1 (and so, due to the formula in Table 3.2, of the related eigenvalue λ_j of the Laplace operator). For example, the lower bound in (3.52) shows that the improvement of reliability constant $\alpha^{-1} = \mu_1^{-1} \approx 2 \times 10^6$ for the Helmholtz equation on the unit square domain with frequency $\omega = 7$ by the factor two requires at least an approximation error

$$(\mu_{h,1}-\mu_1)\mu_{h,1}^{-1}\mu_1^{-1} \approx (\mu_{h,1}-\mu_1)\mu_1^{-2} \leq \alpha^{-1}/2 = \mu_1^{-1}/2.$$

In other words, the relative error in the eigenvalue approximation must satisfy

$$(\mu_{h,1}-\mu_1)\mu_1^{-1} \leq 1/2. \tag{3.53}$$

The computation for $k=1$, ndof = 263169, and $\omega = 7$ in Experiment 1 results in $(\mu_{h,1}-\mu_1)/\mu_1 = 44.1$. Providing $\mu_{h,1}-\mu_1 = \mathcal{O}(\text{ndof}^{-1})$ (cf. Figure 3.3), the approximation error in (3.53) requires about ndof $= 263169 \times 100$ degrees of freedom. Thus, the computational effort for a small improvement of the reliability constant is tremendous.

Huge reliability constants $C(X_h)$ for $\alpha \ll 1$ raise the question whether the improved GUBs result in huge overestimations. Figure 3.5 shows that the answer is in general no.

More precisely, Figure 3.5 displays the error $\|\mathfrak{u}-\mathfrak{u}_h\|_X$ and the GUB $(C(X_h)LS(f;\mathfrak{u}_h))^{1/2}$ of the LSFEM for the Helmholtz equation with frequency $\omega = 4$ and exact solution $\mathfrak{u} = (\phi_1, \nabla\phi_1)$ with $\phi_1(x,y) = \sin(\pi x)\sin(\pi y)$ (left-hand side) and with frequency $\omega = 7$ and exact solution $\mathfrak{u} = (\phi_2, \nabla\phi_2)$ with $\phi_2(x,y) = \sin(2\pi x)\sin(\pi y)$ (right-hand side). The reliability constant $C(X_h)$ results from the three-stage algorithm with $N = 11$. Although $C(X_h)$ is large, the efficiency index $(C(X_h)LS(f;\mathfrak{u}_h))^{1/2}/\|\mathfrak{u}-\mathfrak{u}_h\|_X$ is close to one. For example $(C(X_h)LS(f;\mathfrak{u}_h))^{1/2}/\|\mathfrak{u}-\mathfrak{u}_h\|_X = 1.03$ for $k = 1$, ndof $= \dim X_h = 4225$, $\omega = 4$ and $(C(X_h)LS(f;\mathfrak{u}_h))^{1/2}/\|\mathfrak{u}-\mathfrak{u}_h\|_X = 1.001$ for $k = 2$, ndof $= 14593$, $\omega = 7$.

A more detailed look at the experiment in Figure 3.5 reveals that the regime where the error does not decrease, corresponds to the regime where $\alpha^{-1} \leq (\mu_{h,1} - \mu_1)\mu_1^{-2}$ (or equivalently $1 \leq (\mu_{h,1}-\mu_1)\mu_1^{-1}$). In this regime neither the error nor the residual $LS(f;\mathfrak{u}_h)$ decrease. Further (not displayed) computations suggest that this regime is the same for experiments with different right-hand sides (for example $f \equiv 1$). As the computation overcomes this regime, the error and the residual decrease. Since the error converges asymptotically towards the residual (which is smaller than the error in all computations), a second regime with faster convergence of the error occurs, that is, after the regime without convergence there exists a regime where the error converges faster than the residual. The reliability constant $C(X_h)$, which is close to α^{-1} in the first regime, decreases in the second regime and so still allows for an efficient estimate.

The numerical experiment indicates that the LSFEM works well, if and only if the space X_h allows for good a approximation of the smallest eigenvalue(s). In other words, if $C(X_h) \approx \alpha^{-1}$, the LSFEM will not result in a good approximation. This indicates significant difficulties of the LSFEM for problems with frequencies close to a resonance.

Remark 3.2.9 (Numerical difficulties)**.** *The computation for $k = 3$,* ndof $= 493057$*, and $\omega = 7$ results in $C(X_h)LS(f;\mathfrak{u}_h) < \|\mathfrak{u}-\mathfrak{u}_h\|_X^2$. Since $C(X_h)LS(f;\mathfrak{u}_h)$ is an upper bound for $\|\mathfrak{u}-\mathfrak{u}_h\|_X^2$, this indicates numerical difficulties. These difficulties might result from*

- *an inexact solve of the linear problem $\mathfrak{u}_h = \arg\min_{x_h\in X_h} LS(f;x_h)$ (thus, the Galerkin orthogonality, which is an important assumption in the proof of Theorem 3.1.4, does not hold) or*
- *an inexact solve of the eigenvalue problem, which results in lower bounds for the discrete eigenvalues $\mu_{h,k}$ with $k = 1,\dots,N$ and so leads to $C(X_h) < \alpha(X_h)$ with $\alpha(X_h)$ from (3.19).*

Experiment 3 (Estimator competition for Helmholtz). This experiment approximates the solution $\mathfrak{u} = \arg\min_{x\in X} LS(f;x)$ to the Helmholtz equation with constant right-hand side $f \equiv 1$ and frequency $\omega = 4$ on the L-shaped domain $\Omega = (-1,1)^2 \setminus [0,1)^2$ by the solution $\mathfrak{u}_h = (u_h, p_h) = \arg\min_{x_h\in X_h} LS(f;x_h)$ to the LSFEM with $X_h = S_0^k(\mathcal{T}) \times RT_{k-1}(\mathcal{T})$ and $k = 1,2,3$. It utilizes the adaptive mesh refinement from Algorithm 3 on page 137 with bulk parameter $\Theta = 0.3$ and refinement indicator

$$\eta^2(T) = \|\nabla u_h - p_h\|_{L^2(T)}^2 + \|\omega^2 u_h + \operatorname{div} p_h + f\|_{L^2(T)}^2 \qquad \text{for all } T \in \mathcal{T}.$$

The computation stops, if the error

$$\|\mathfrak{u}-\mathfrak{u}_h\|_X \leq \text{tol} \qquad \text{with tol} = 10^{-n} \text{ and } n = 0,\dots,3. \tag{3.54}$$

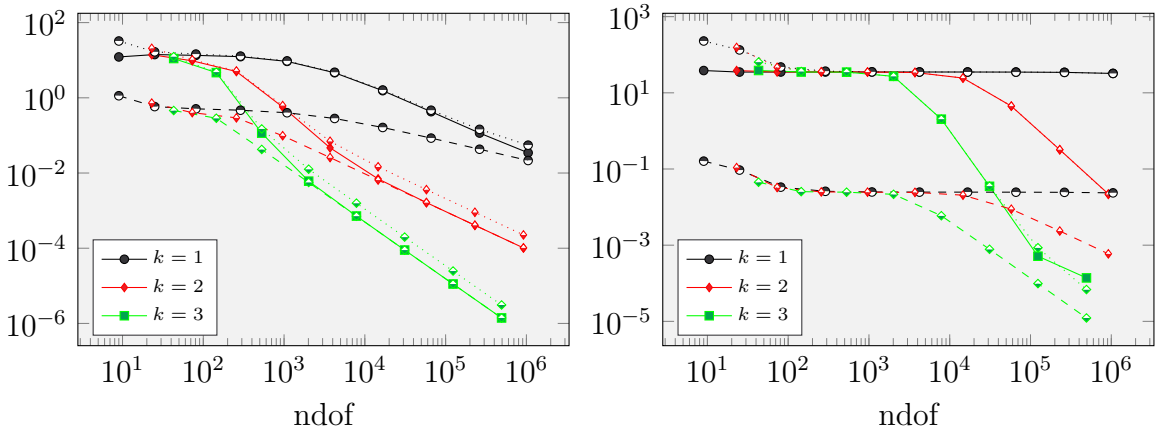
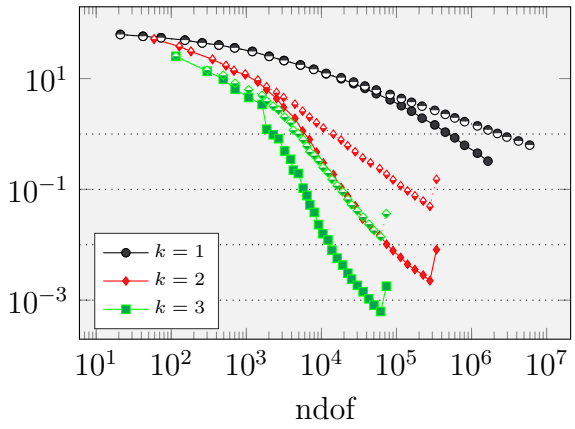


Figure 3.5: Error $\|\mathfrak{u}-\mathfrak{u}_h\|_X$ (solid line), GUB $(C(X_h)LS(f;\mathfrak{u}_h))^{1/2}$ (dotted line), and residual $LS(f;\mathfrak{u}_h)$ (dashed line) with $N = 11$ and polynomial degree $k = 1, 2, 3$ for the Helmholtz equation with $\omega = 4$ (left) and $\omega = 7$ (right) in Experiment 2



		tol			
k	GUB	10^0	10^{-1}	10^{-2}	10^{-3}
1	a)	195 s	–	–	–
1	b)	334 s	–	–	–
2	a)	2 s	6 s	25 s	–
2	b)	1 s	11 s	–	–
3	a)	1 s	3 s	6 s	19 s
3	b)	1 s	2 s	–	–

Figure 3.6: Convergence history plot of the improved GUB a) (filled markers) and the natural GUB b) (half-filled markers) with polynomial degree $k = 1, 2, 3$ and the time in seconds of the adaptive algorithm (including the mesh refinements and the computation of the solutions $\mathfrak{u}_h$ and GUBs) with stopping criterion (3.54) for tol $= 10^0, 10^{-1}, 10^{-2}, 10^{-3}$ in Experiment 3

To guarantee (3.54), the algorithm utilizes a) the improved GUB $(C(X_h)LS(f;\mathfrak{u}_h))^{1/2}$ with $C(X_h)$ from the three-stage algorithm on page 23 with $N = 7$ and exact eigenvalues $\mu_1 = \mu_1^{\text{low}}, \dots, \mu_8 = \mu_8^{\text{low}}$ in Stage 1 (the exact values $\mu_1, \dots, \mu_8$ result from the formula in Table 3.2 and the accurate approximations of the Dirchlet eigenvalues of the Laplace operators $\lambda_1, \dots, \lambda_8$ in [TB06] and [Gal14a]) and b) the natural GUB $(\alpha^{-1}LS(f;\mathfrak{u}_h))^{1/2}$. The convergence history plot in Figure 3.6 displays both GUBs and shows that the improved GUB a) allows for the stopping criterion (3.54) with less iterations in the adaptive algorithm. This results in a much faster computation (see Figure 3.6). Since the computation of $\mathfrak{u}_h = \arg\min_{x_h \in X_h} LS(f;x_h)$ runs out of memory (for $k = 1$ and $\text{ndof} = \dim X_h > 5949710$) or the huge condition number of the system matrices results in a large error with the direct solver MUMPS (for $k = 2$ and $\text{ndof} > 341024$ and $k = 3$ and $\text{ndof} > 61573$), the improved GUB a) allows for much smaller tolerances.

Experiment 4 (Estimator competition for Maxwell). This experiment investigates GUBs for the Maxwell equations with frequency $\omega = 1$ and Fichera corner domain $\Omega = (-1,1)^3 \setminus [0,1)^3$. The experiment utilizes the Nédélec finite element space of first kind, which reads, for regular triangulations $\mathcal{T}$ of the bounded Lipschitz domain $\Omega \subset \mathbb{R}^3$, polynomial degree $k \in \mathbb{N}_0$, and $\mathcal{N}^k(T) := \mathbb{P}_k(T;\mathbb{R}^3) + \mathbb{P}_k(T;\mathbb{R}^3) \times \mathrm{id}$ with the identity map $\mathrm{id} : T \to T$ for all $T \in \mathcal{T}$,

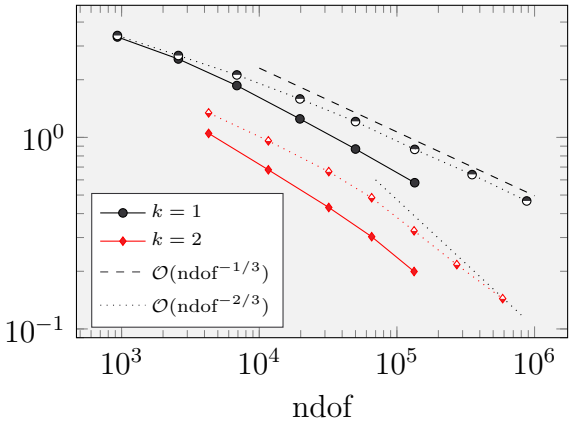
$$\begin{aligned} \mathcal{N}^k(\mathcal{T}) &:= \{v \in H(\operatorname{curl},\Omega) \mid v|_T \in \mathcal{N}^k(T) \text{ for all } T \in \mathcal{T}\}, \\ \mathcal{N}_0^k(\mathcal{T}) &:= H_0(\operatorname{curl},\Omega) \cap \mathcal{N}^k(\mathcal{T}). \end{aligned} \tag{3.55}$$

Set the discrete space $X_h := \mathcal{N}_0^k(\mathcal{T}) \times \mathcal{N}^k(\mathcal{T})$ for $k = 0, 1$ and regular triangulations $\mathcal{T}$ of the domain Ω into tetrahedra. The exact eigenvalues $0 < \lambda_1 \le \lambda_2 \le \dots$ of the curl curl operator are unknown. The lower eigenvalue bounds $\mu_1^{\text{low}} \le \mu_1, \dots, \mu_{15}^{\text{low}} \le \mu_{15}$ result from the formula in Table 3.2 and the guaranteed lower and upper eigenvalue bounds for $\lambda_1, \dots, \lambda_{15}$ from [BBB17, Sec. 6.2]. The lower eigenvalue bounds lead to the approximation $\alpha^{-1} := 1/\mu_1^{\text{low}} = 7.01$. Figure 3.7 displays the convergence history plot of the GUBs a) $(C(X_h)LS(f;\mathfrak{u}_h))^{1/2}$ (with $C(X_h)$ from the three-stage algorithm with $N = 14$) and b) $(\alpha^{-1}LS(f;\mathfrak{u}_h))^{1/2}$ with discrete solution $\mathfrak{u}_h = (u_h, \sigma_h) = \arg\min_{x_h \in X_h} LS(f;x_h)$ to the LSFEM for the Maxwell equation with right-hand side $f \equiv (1,1,1)^\top$ and frequency $\omega = 1$. The adaptively refined meshes result from Algorithm 3 on page 137 with bulk parameter $\Theta = 0.3$ and refinement indicator

$$\eta^2(T) := \|\sigma_h - \operatorname{curl} u_h\|_{L^2(T)}^2 + \|\operatorname{curl} \sigma_h - \omega^2 u_h - f\|_{L^2(T)}^2 \qquad \text{for all } T \in \mathcal{T}.$$

The natural GUB b) is about 1.5 times bigger than the improved GUB a) on fine meshes (see the tabular in Figure 3.7). Since the experiment suggests $LS(f;\mathfrak{u}_h)^{1/2} = \mathcal{O}(\text{ndof}^{-1/3})$ for $k = 0$ and $\text{ndof} = \dim X_h$, the adaptive algorithm with stopping criterion (3.54) and GUB b) stops one or two iterations after the adaptive algorithm with GUB a). In contrast to the previous experiment, this does not lead to a faster computation. For example, the algorithm with $\text{tol} = 1$ and $k = 0$ requires 121 seconds and $\text{ndof} = 50014$ with GUB a) and 46 seconds and $\text{ndof} = 134628$ with GUB b). Similar results hold for $k = 1$.

Discussion. The overall conclusions from all the numerical benchmarks reported in this and the previous section are in agreement with the theoretical predictions of this work. The



$k = 0$		$k = 1$	
ndof	$C(X_h)$	ndof	$C(X_h)$
932	6.77	4288	4.28
2584	6.44	11640	3.49
6874	5.43	31992	2.96
19748	4.33	65572	2.75
50014	3.62	133252	2.63
134628	3.14	–	–

Figure 3.7: Convergence history plot for the GUB a) (solid line) and b) (dotted line) with $\omega = 1$ and polynomial degree $k = 0, 1$ and the reliability constant $C(X_h)$ in Experiment 4

improvement of the reliability constant $C(X_h)$ is visible in all experiments and moderate for the Poisson model problem without degenerated geometry but can exceed several orders of magnitude for certain parameters of ω in the Helmholtz and Maxwell equations. A possible explanation starts with the overall observation that $\|\mathfrak{u} - \mathfrak{u}_h\|_X^2 \leq LS(f; \mathfrak{u}_h)$ in all experiments so that $1 \leq C(X_h) \leq \alpha^{-1}$ and α^{-1} moderate merely implies a moderate improvement of $\alpha C(X_h) \leq 1$. For critical parameter $2 \leq \omega \leq 10$ in Experiment 1, Table 3.3 displays $5 \leq \alpha^{-1} \leq 1184260$ and allows for a dramatic improvement of $\alpha C(\mathcal{T}) \ll 1$. In those examples, a few eigenfunctions need to be resolved so that (3.24) leads to $C(X_h)$ close to $\mu(m+1)^{-1} \ll \mu(1)^{-1} = \alpha^{-1}$ with a moderate $m \in \mathbb{N}$. This reduction factor of nearly $\alpha\mu(m+1)^{-1}$ for fine meshes has to be evaluated in relation to the additional costs for several eigenvalue calculations.

The remaining part of this discussion focuses on the guaranteed error control as a stopping criterion of an adaptive mesh refinement with the guaranteed error control (3.54), that is, the adaptive algorithm stops if the error $\|\mathfrak{u} - \mathfrak{u}_h\|_X$ is smaller than a given tolerance $\text{tol} > 0$. Suppose that a fine triangulation $\mathcal{T}$ satisfies $C(X_h)LS(f; \mathfrak{u}_h) \leq \text{tol}^2$ with ndof degrees of freedom in the discrete system. For a simplified comparison, suppose that the computational costs CPU are proportional to ndof (for an optimal iterative solver despite the fact that our numerical examples run with the direct solver MUMPS in FEniCS). In all numerical experiments the time $t(C(X_h))$ for computing $C(X_h)$ is linear to the time $t(\mathfrak{u}_h)$ for solving the LSFEM $\mathfrak{u}_h = \arg\min_{x_h \in X_h} LS(f; x_h)$, that is $t(C(X_h)) = \rho(N)t(\mathfrak{u}_h)$ with $\rho(N) > 0$ ($\rho(7) = 4$ in Experiment 2 and $\rho(14) = 20$ in Experiment 3 of this section). This suggests that the adaptive algorithm may stop with the triangulation $\mathcal{T}$, but requires extra costs of $\rho(N)$ CPU for the more expansive improved GUB $C(X_h)LS(f; \mathfrak{u}_h)$. In the present model situations, the usage of the GUB $\alpha^{-1}LS(f; \mathfrak{u}_h)$ implies further mesh refinements until the bound $\alpha^{-1}LS(f; \mathfrak{u}'_h) \leq \text{tol}^2$ holds for a discrete solution $\mathfrak{u}'_h$ with respect to a much finer mesh $\mathcal{T}'$ with ndof' degrees of freedom. In the case of an optimal convergence $LS(f; \mathfrak{u}_h) = \mathcal{O}(\text{ndof}^{-k})$ of the adaptive algorithm in 2D with polynomial degree $k \in \mathbb{N}$ of X_h, the stopping criterion $C(X_h)LS(f; \mathfrak{u}_h) = \text{tol}^2 = \alpha^{-1}LS(f; \mathfrak{u}'_h)$ results in $(C(X_h)\alpha)^{1/k} = \text{ndof}/\text{ndof}'$. Hence, if $C(X_h)\alpha)^{1/k} \leq 1/\rho(N)$, the three-stage algorithm

from Section 3.1.2 appears less expensive in the computational online costs. This calculation leaves out the additional mesh refinements required in the adaptive algorithm to compute $\mathcal{T}'$ and therefore is very conservative. This discussion also ignores the fact that the reliability constant may be computed with discrete space $X_H \subset X_h$ of moderate size $\dim X_H$ and may utilize $C(X_H) \leq C(X_h)$ for all refinements $\mathcal{T}$. Since the three-stage algorithm with $N \in \mathbb{N}$ and small discrete space X_H with large polynomial degree $k \in \mathbb{N}$ often allows for $C(X_H)$ close to $\mu(N+1)^{-1} \ll \alpha^{-1}$, this ansatz is very advantageous for higher polynomial degrees k.

Based on this discussion, the proposed algorithm appears advantageous if $\alpha\rho(N)^k < \mu(m+1)$ for moderate m and sufficiently small tolerances in guaranteed error control. If $\mu(m+1) < \alpha\rho(N)^k$, the computation of $C(X_H)$ on a smaller space $X_H \subset X_h$ can result in significantly smaller GUBs and so allows for smaller tolerances tol and faster computations.

3.2.3 Stokes

The Stokes equation describes the flow of incompressible fluids with large viscosities. The survey article [BG98] emphasizes the practical relevance of the LSFEM for the Stokes (and the related Navier-Stokes) problem and the huge scientific interest on this topic. However, to the author's knowledge, there exist no results on the asymptotic behaviour of the residual $LS(f;\mathfrak{u}_h)$. Furthermore, explicit values for the equivalence constants α and β with (3.1) are missing. This section investigates both. It focuses on the LSFEM formulation from [CLW04, CW07, BC17a, BC17b], which reads as follows. Let $\Omega \subset \mathbb{R}^d$ with $2 \leq d \in \mathbb{N}$ be a bounded Lipschitz domain, $f \in L^2(\Omega;\mathbb{R}^d)$ a given external body force, and $\gamma > 0$ an arbitrary weight. The solution

$$(u,\sigma) \in X := H^1_0(\Omega;\mathbb{R}^d) \times \Sigma \text{ with } \Sigma := \{\tau \in H(\operatorname{div},\Omega;\mathbb{R}^{d\times d}) \mid \int_\Omega \operatorname{tr}(\tau)\,\mathrm{d}x = 0\} \tag{3.56}$$

to the pseudostress-velocity formulation (2.23) minimizes the least-squares functional

$$LS(f;v,\tau) := \|\operatorname{dev}\tau - \nabla v\|^2_{L^2(\Omega)} + \gamma\|f + \operatorname{div}\tau\|^2_{L^2(\Omega)} \qquad \text{over all } (v,\tau) \in X. \tag{3.57}$$

Theorem 2.5.1 proves that Σ is a Hilbert space with inner product $(\bullet,\bullet)_\Sigma$ and induced norm $\|\bullet\|^2_\Sigma = \|\operatorname{dev}\bullet\|^2_{L^2(\Omega)} + \|\operatorname{div}\bullet\|^2_{L^2(\Omega)}$.

Theorem 3.2.10 (Stokes eigenvalue problem)**.** *There exist countably many eigenvalues* $0 < \lambda_1 \leq \lambda_2 \leq \dots$ *with eigenfunctions* $\phi_k \in \Sigma \setminus \{0\}$ *and*

$$(\operatorname{div}\phi_k, \operatorname{div}\tau)_{L^2(\Omega)} = \lambda_k\,(\operatorname{dev}\phi_k, \operatorname{dev}\tau)_{L^2(\Omega)} \qquad \textit{for all } \tau \in \Sigma \textit{ and } k \in \mathbb{N}. \tag{3.58}$$

The eigenvalues $\lambda_k \to \infty$ *as* $k \to \infty$ *and the eigenfunctions* $(\operatorname{dev}\phi_k, \operatorname{dev}\phi_\ell)_{L^2(\Omega)} = \delta_{k\ell}$ *are orthonormal for all* $k,\ell \in \mathbb{N}$*. The space* Σ *decomposes with* $\Sigma_0 := \{\tau \in \Sigma \mid \operatorname{div}\tau = 0\}$*,* $\Sigma_1 := \{pI_{d\times d} \mid p \in H^1(\Omega)$ *and* $\int_\Omega p\,\mathrm{d}x = 0\}$*, and the closure* Σ_2 *of* $\operatorname{span}\{\phi_k \mid k \in \mathbb{N}\}$ *with respect to the norm* $\|\bullet\|_\Sigma$ *into*

$$\Sigma = \Sigma_0 \oplus \Sigma_1 \oplus \Sigma_2. \tag{3.59}$$

The split is orthogonal with respect to the inner product $(\bullet,\bullet)_\Sigma$*, that is* $(\tau_k,\tau_\ell)_\Sigma = 0$ *for all* $\tau_k \in \Sigma_k$ *and* $\tau_\ell \in \Sigma_\ell$ *with* $k,\ell = 1,2,3$ *and* $k \neq \ell$*.*

Proof. This theorem follows from [MMR15, Thm. 3.5]. □

Remark 3.2.11 (Special case of [MMR15])**.** *The boundary in [MMR15] decomposes (in the notation from [MMR15]) into $\partial\Omega = \Gamma \cup \Sigma$ with disjoint components Γ and Σ. Theorem 3.2.10 utilizes the special case $\partial\Omega = \Gamma$. The proof of this simpler special case requires the replacement of the space $\mathcal{W} = H(\operatorname{div}, \Omega; \mathbb{R}^{d\times d})$ by $\{\tau \in \mathcal{W} \mid \int_\Omega \operatorname{tr}(\tau)\,\mathrm{d}x = 0\} \subset \mathcal{W}$ and the application of Theorem 2.5.1 instead of [MMR15, Lem. 2.3].*

Remark 3.2.12 (Related eigenvalue problem)**.** *The eigenvalues in Theorem 3.2.10 equal [MMR15] the eigenvalues to the Stokes eigenvalue problem: Seek an eigenvalue $\lambda \in \mathbb{R}$ and an eigenfunction $0 \neq (u, p) \in H^1_0(\Omega; \mathbb{R}^d) \times H^1(\Omega)$ with*

$$-\operatorname{div}(\nabla u) + \nabla p = \lambda u, \qquad \operatorname{div} u = 0, \qquad \text{and} \qquad \int_\Omega p\,\mathrm{d}x = 0.$$

Lemma 3.2.13 (Deviator)**.** *The deviator satisfies, for all $\tau, \vartheta \in L^2(\Omega; \mathbb{R}^{d\times d})$,*

$$(\operatorname{dev}\tau, \operatorname{dev}\vartheta)_{L^2(\Omega)} = (\operatorname{dev}\tau, \vartheta)_{L^2(\Omega)} = (\tau, \operatorname{dev}\vartheta)_{L^2(\Omega)}, \tag{3.60a}$$

$$\|\tau\|^2_{L^2(\Omega)} = \|\operatorname{dev}\tau\|^2_{L^2(\Omega)} + d^{-1}\|\operatorname{tr}(\tau)\|^2_{L^2(\Omega)}. \tag{3.60b}$$

Proof. The definition $\operatorname{dev}\tau := \tau - d^{-1}\operatorname{tr}(\tau) I_{d\times d}$ for all $\tau \in L^2(\Omega; \mathbb{R}^{d\times d})$ and simple calculations imply these properties. □

The properties of the solution $\mathfrak{u} \in X$ to (2.23) imply $LS(f; v, \tau) = \|\mathfrak{u} - (v, \tau)\|^2_a$ with squared norm $\|(v,\tau)\|^2_a = \|\operatorname{dev}\tau - \nabla v\|^2_{L^2(\Omega)} + \gamma\|\operatorname{div}\tau\|^2_{L^2(\Omega)}$ for all $(v, \tau) \in X$. The norm $\|\bullet\|_a$ is equivalent to the norm $\|(v,\tau)\|_X := (\|\nabla v\|^2_{L^2(\Omega)} + \|\operatorname{dev}\tau\|^2_{L^2(\Omega)} + \gamma\|\operatorname{div}\tau\|^2_{L^2(\Omega)})^{1/2}$ for all (v, τ) in the Hilbert space X [CLW04, Thm. 4.2]. Let the inner product $a(\bullet, \bullet)$ induce $\|\bullet\|_a$ and let $(\bullet, \bullet)_X$ induce $\|\bullet\|_X$.

Reduced ansatz space

Since the solution $(u, \sigma) \in X := H^1_0(\Omega; \mathbb{R}^d) \times \Sigma$ to the pseudostress-velocity formulation (2.23) of the Stokes problem satisfies $\operatorname{dev}\sigma = \nabla u$, the divergence $\operatorname{div} u = \operatorname{tr}(\nabla u) = \operatorname{tr}(\operatorname{dev}\sigma) = 0$. In other words, the velocity field

$$u \in Z := \{v \in H^1_0(\Omega; \mathbb{R}^d) \mid \operatorname{div} v = 0\} \subset H^1_0(\Omega; \mathbb{R}^d). \tag{3.61}$$

This motivates least-squares schemes with reduced ansatz space

$$X_{\mathrm{red}} := Z \times \Sigma \tag{3.62}$$

and discrete subspaces $X_{\mathrm{red},h} \subset X_{\mathrm{red}}$ in the sense that the discrete minimizer $\mathfrak{u}_h = (u_h, \sigma_h) = \arg\min_{x_h \in X_{\mathrm{red,h}}} LS(f; x_h)$ of the least-squares functional $LS(f; \bullet)$ from (3.57) approximates the solution $\mathfrak{u} = (u, \sigma) \in X_{\mathrm{red}} \subset X$ to the Stokes problem (2.23).

Lemma 3.2.14 (Orthogonal system in Z)**.** *The linear hull of $\{\operatorname{div}\phi_k \mid k \in \mathbb{N}\}$ with the eigenfunctions ϕ_k from Theorem 3.2.10 is dense in Z, that is*

$$Z = \overline{\operatorname{span}\{\operatorname{div}\phi_k \mid k \in \mathbb{N}\}}^{\|\nabla\bullet\|_{L^2(\Omega)}}.$$

Proof. Step 1 ("⊇"). Let $k \in \mathbb{N}$. Lemma 3.2.13, (3.58), $\operatorname{div} I_{d\times d} = 0$, and $\operatorname{dev} I_{d\times d} = 0$ imply

$$(\operatorname{div}\phi_k, \operatorname{div}\tau)_{L^2(\Omega)} = \lambda_k\,(\operatorname{dev}\phi_k, \tau)_{L^2(\Omega)} \qquad \text{for all } \tau \in H(\operatorname{div}, \Omega; \mathbb{R}^{d\times d}).$$

In other words, $\operatorname{div}\phi_k \in H^1_0(\Omega;\mathbb{R}^d)$ with $\nabla\operatorname{div}\phi_k = -\lambda_k \operatorname{dev}\phi_k$. Since $0 = -\lambda_k \operatorname{tr}(\operatorname{dev}\phi_k) = \operatorname{tr}(\nabla\operatorname{div}\phi_k) = \operatorname{div}\operatorname{div}\phi_k$, the function $\operatorname{div}\phi_k$ is an element in the closed space Z.

Step 2 ("⊆"). Let $z \in Z$. The combination of Theorem 2.5.2 and Theorem 2.2.3 results in the surjectivity of the divergence operator $\operatorname{div} : \Sigma_1 \oplus \Sigma_2 \to L^2(\Omega;\mathbb{R}^d)$. Since $Z \subset L^2(\Omega;\mathbb{R}^d)$, the surjectivity proves the existence of a function $\tau \in \Sigma_1 \oplus \Sigma_2$ with $\operatorname{div}\tau = z$ and decomposition $\tau = pI_{d\times d} + \tau_2$ with $p \in H^1(\Omega)$, $\int_\Omega p\,\mathrm{d}x = 0$, and $\tau_2 \in \Sigma_2$. Since $\operatorname{div} z = 0$, integration by parts and $z \in Z \subset H^1_0(\Omega;\mathbb{R}^d) \subset H_0(\operatorname{div},\Omega)$ reveal $(z, \nabla p)_{L^2(\Omega)} = -(\operatorname{div} z, p)_{L^2(\Omega)} = 0$. This identity, $\operatorname{div}(pI_{d\times d}) = \nabla p$, $z = \operatorname{div}(pI_{d\times d} + \tau_2)$, and the orthogonality of Σ_1 and Σ_2 with respect to $(\bullet,\bullet)_\Sigma$ result in

$$(\nabla p, \nabla p)_{L^2(\Omega)} = (\operatorname{div} pI_{d\times d}, \nabla p)_{L^2(\Omega)} = -(\operatorname{div}\tau_2, \nabla p)_{L^2(\Omega)} = -(\tau_2, pI_{d\times d})_\Sigma = 0.$$

The combination with $\int_\Omega p\,\mathrm{d}x = 0$ shows $p = 0$ and so $z = \operatorname{div}\tau_2$. Since Σ_2 is the closure of $\operatorname{span}\{\phi_k \mid k \in \mathbb{N}\}$ with respect to the norm $\|\bullet\|_\Sigma$, there exist coefficients $\tau_{2,k} \in \mathbb{R}$ with $\tau_2 = \sum_{k=1}^\infty \tau_{2,k}\phi_k$. Since the eigenfunctions $\phi_1, \phi_2, \dots$ are orthonormal, that is $(\operatorname{dev}\phi_k, \operatorname{dev}\phi_\ell)_{L^2(\Omega)} = \delta_{k\ell}$ for all $k, \ell \in \mathbb{N}$, and satisfy $\nabla\operatorname{div}\phi_k = -\lambda_k\operatorname{dev}\phi_k$ (see Step 1) for all $k \in \mathbb{N}$, it holds

$$\begin{aligned}\sum_{k=1}^\infty \tau_{2,k}^2\lambda_k^2 &= \sum_{k=1}^\infty \tau_{2,k}^2 \|\lambda_k \operatorname{dev}\phi_k\|^2_{L^2(\Omega)} = \|\sum_{k=1}^\infty \tau_{2,k}\lambda_k\operatorname{dev}\phi_k\|^2_{L^2(\Omega)}\\ &= \|\sum_{k=1}^\infty \tau_{2,k}\nabla\operatorname{div}\phi_k\|^2_{L^2(\Omega)} = \|\nabla\operatorname{div}\tau_2\|^2_{L^2(\Omega)} = \|\nabla z\|^2_{L^2(\Omega)} < \infty.\end{aligned}$$

Thus, $\|\nabla(z - \sum_{k=1}^n \tau_{2,k}\operatorname{div}\phi_k)\|^2_{L^2(\Omega)} = \|\sum_{k=n+1}^\infty \tau_{2,k}\nabla\operatorname{div}\phi_k\|_{L^2(\Omega)}{}^2 = \sum_{k=n+1}^\infty \tau_{2,k}^2\lambda_k^2 \to 0$ as $n \to \infty$ and so

$$z = \operatorname{div}\tau_2 = \sum_{k=1}^\infty \tau_{2,k}\operatorname{div}\phi_k \in \overline{\operatorname{span}\{\operatorname{div}\phi_k \mid k\in\mathbb{N}\}}^{\|\nabla\bullet\|_{L^2(\Omega)}}. \qquad \square$$

Let $\gamma > 0$ be an arbitrary weight and let $a(\bullet,\bullet)$ and $(\bullet,\bullet)_X$ induce the norms

$$\begin{aligned}\|(v,\tau)\|_a &:= LS(0;v,\tau)^{1/2} = (\|\operatorname{dev}\tau - \nabla v\|^2_{L^2(\Omega)} + \gamma\|\operatorname{div}\tau\|^2_{L^2(\Omega)})^{1/2}, \qquad (3.63)\\ \|(v,\tau)\|_X &:= (\|\operatorname{dev}\tau\|^2_{L^2(\Omega)} + \|\nabla v\|^2_{L^2(\Omega)} + \gamma\|\operatorname{div}\tau\|^2_{L^2(\Omega)})^{1/2} \qquad \text{for all } (v,\tau) \in X_{\mathrm{red}}.\end{aligned}$$

Recall the eigenpairs $(\lambda_k, \phi_k) \in \mathbb{R}\times\Sigma\setminus\{0\}$ from Theorem 3.2.10 for all $k \in \mathbb{N}$ and set

$$\mu_0 := 1 \quad\text{and}\quad \psi_0 \in \{0\}\times(\Sigma_0\oplus\Sigma_1) \subset X_{\mathrm{red}}, \qquad (3.64a)$$

$$\mu_{2k-1} := 1 - (\gamma\lambda_k + 1)^{-1/2} \quad\text{and}\quad \psi_{2k-1} := ((1+\gamma\lambda_k)^{1/2}/\lambda_k\,\phi_k, -\operatorname{div}\phi_k) \in X_{\mathrm{red}}, \qquad (3.64b)$$

$$\mu_{2k} := 1 + (\gamma\lambda_k + 1)^{-1/2} \quad\text{and}\quad \psi_{2k} := ((1+\gamma\lambda_k)^{1/2}/\lambda_k\,\phi_k, \operatorname{div}\phi_k) \in X_{\mathrm{red}}. \qquad (3.64c)$$

Theorem 3.2.15 (Eigenvalues of $a(\bullet,\bullet)$ and $(\bullet,\bullet)_X$)**.** *The formulae in (3.64) define eigenpairs with*

$$a(\psi_k, x) = \mu_k\,(\psi_k, x)_X \qquad \textit{for all } x \in X_{\mathrm{red}} \textit{ and } k \in \mathbb{N}_0. \qquad (3.65)$$

Proof. Step 1 (Decomposition of the inner products). Let $(v,\tau),(w,\vartheta) \in X_{\mathrm{red}}$. Theorem 3.2.10 and Lemma 3.2.14 imply the existence of functions $\tau_0,\vartheta_0 \in \Sigma_0$ and $\tau_1,\vartheta_1 \in \Sigma_1$ as well as the existence of coefficients $\tau_{2,k},\vartheta_{2,k},v_k,w_k \in \mathbb{R}$ for all $k \in \mathbb{N}$ with

$$v = \sum_{k\in\mathbb{N}} v_k \mathrm{div}\,\phi_k, \quad w = \sum_{k\in\mathbb{N}} w_k \mathrm{div}\,\phi_k, \quad \tau = \tau_0 + \tau_1 + \sum_{k\in\mathbb{N}} \tau_{2,k}\phi_k, \quad \vartheta = \vartheta_0 + \vartheta_1 + \sum_{k\in\mathbb{N}} \vartheta_{2,k}\phi_k.$$

The orthogonality of the normed eigenfunctions ϕ_k and $\tau_0,\tau_1,\vartheta_0,\vartheta_1$ and the identity $\nabla \mathrm{div}\,\phi_k = -\lambda_k \mathrm{dev}\,\phi_k$ for all $k \in \mathbb{N}$ allow for the formal calculation

$$\begin{aligned}
a(v,\tau;w,\vartheta) &= (\mathrm{dev}\,\tau - \nabla v, \mathrm{dev}\,\vartheta - \nabla w)_{L^2(\Omega)} + \gamma\,(\mathrm{div}\,\tau, \mathrm{div}\,\vartheta)_{L^2(\Omega)} \\
&= \sum_{k\in\mathbb{N}} ((\tau_{2,k} + \lambda_k v_k)\mathrm{dev}\,\phi_k, (\vartheta_{2,k} + \lambda_k w_k)\mathrm{dev}\,\phi_k)_{L^2(\Omega)} + (\mathrm{dev}\,\tau_0, \mathrm{dev}\,\vartheta_0)_{L^2(\Omega)} \\
&\quad + \sum_{k\in\mathbb{N}} \gamma\lambda_k\,(\tau_{2,k}\mathrm{dev}\,\phi_k, \vartheta_{2,k}\mathrm{dev}\,\phi_k)_{L^2(\Omega)} + \gamma\,(\mathrm{div}\,\tau_1, \mathrm{div}\,\vartheta_1)_{L^2(\Omega)} \\
&= \sum_{k\in\mathbb{N}} \begin{pmatrix} \tau_{2,k} \\ v_k \end{pmatrix} \cdot \begin{pmatrix} 1+\gamma\lambda_k & \lambda_k \\ \lambda_k & \lambda_k^2 \end{pmatrix} \begin{pmatrix} \vartheta_{2,k} \\ w_k \end{pmatrix} + (\mathrm{dev}\,\tau_0, \mathrm{dev}\,\vartheta_0)_{L^2(\Omega)} + \gamma\,(\mathrm{div}\,\tau_1, \mathrm{div}\,\vartheta_1)_{L^2(\Omega)}.
\end{aligned}$$

Similar arguments imply

$$\begin{aligned}
(v,\tau;w,\vartheta)_X = &\sum_{k\in\mathbb{N}} \begin{pmatrix} \tau_{2,k} \\ v_k \end{pmatrix} \cdot \begin{pmatrix} 1+\gamma\lambda_k & 0 \\ 0 & \lambda_k^2 \end{pmatrix} \begin{pmatrix} \vartheta_{2,k} \\ w_k \end{pmatrix} + (\mathrm{dev}\,\tau_0, \mathrm{dev}\,\vartheta_0)_{L^2(\Omega)} \\
&+ \gamma\,(\mathrm{div}\,\tau_1, \mathrm{div}\,\vartheta_1)_{L^2(\Omega)}.
\end{aligned}$$

Step 2 (Computation of eigenpairs). The decomposition of the inner products in Step 1 shows that $\mu_0 = 1$ and $\psi_0 \in \{0\} \times (\Sigma_0 \oplus \Sigma_1)$ satisfy (3.65). For all $k \in \mathbb{N}$ and $(w,\vartheta) \in X_{\mathrm{red}}$ the decomposition in Step 1 results in

$$\begin{aligned}
a(\psi_{2k-1};w,\vartheta) &= \begin{pmatrix} (1+\gamma\lambda_k)^{1/2}/\lambda_k \\ -1 \end{pmatrix} \cdot \begin{pmatrix} 1+\gamma\lambda_k & \lambda_k \\ \lambda_k & \lambda_k^2 \end{pmatrix} \begin{pmatrix} \vartheta_{2,k} \\ w_k \end{pmatrix} \\
&= \mu_{2k-1} \begin{pmatrix} (1+\gamma\lambda_k)^{1/2}/\lambda_k \\ -1 \end{pmatrix} \cdot \begin{pmatrix} 1+\gamma\lambda_k & 0 \\ 0 & \lambda_k^2 \end{pmatrix} \begin{pmatrix} \vartheta_{2,k} \\ w_k \end{pmatrix} = \mu_{2k-1}\,(\psi_{2k-1};w,\vartheta)_X.
\end{aligned}$$

Analogously, $a(\psi_{2k};w,\vartheta) = \mu_{2k}\,(\psi_{2k};w,\vartheta)_X$ follows for all $k \in \mathbb{N}$ and $(w,\vartheta) \in X_{\mathrm{red}}$. □

Theorem 3.2.16 (Properties of the restricted inner products)**.** *The inner products $a(\bullet,\bullet) : X_{\mathrm{red}} \times X_{\mathrm{red}} \to \mathbb{R}$ and $(\bullet,\bullet)_X : X_{\mathrm{red}} \times X_{\mathrm{red}} \to \mathbb{R}$ from (3.63) with spaces Z from (3.61), Σ from (3.56), and $X_{\mathrm{red}} := Z \times \Sigma$ satisfy the hypotheses* (**H1**)–(**H4**) *on page 15.*

Proof. The arguments from the proof of Theorem 3.2.3 imply this theorem. □

Since Theorem 3.2.16 validates the four hypotheses (**H1**)–(**H4**), the results from Section 3.1.1–3.1.2 apply for any discretization $X_{\mathrm{red},h} = Z_h \times \Sigma_h$ with (**D**). While discretizations and implementations of Σ conforming subspaces Σ_h are well established (cf. [BC05] for an implementation of lowest-order Raviart-Thomas elements in a few lines of Matlab code), the discretization of Z as for example in [AQ92, GN14a, GN14b, NS16, QZ07, SV85, Zha08] is unusual and often not practical in the sense that the spaces require higher polynomial degrees or restrictions on the geometry of the mesh (see [DKS13, Sec. 3.3] for a more detailed discussion).

Full ansatz space

The previous section shows that the reduced ansatz space $X_{\mathrm{red}} \subset X = H^1_0(\Omega;\mathbb{R}^d) \times \Sigma$ with X from (3.56) leads to (**H1**)–(**H4**) for the Stokes problem and so allows for the applications of the results from Section 3.1.1–3.1.2. However, the application of well-established finite elements like Courant elements results in discrete spaces $X_h \subset X$ with $X_h \not\subset X_{\mathrm{red}}$. Therefore, the remainder of this section investigates the LSFEM for the Stokes problem with full ansatz space X. Recall the space Z from (3.61) and set the orthogonal complement

$$Z^\perp := \{z^\perp \in H^1_0(\Omega;\mathbb{R}^d) \mid (\nabla z^\perp, \nabla z)_{L^2(\Omega)} = 0 \text{ for all } z \in Z\}. \tag{3.66}$$

The space $Z^\perp$ allows for the following well-known (see for example [CD15, Eq. 2.4]) alternative characterization of the LBB constant C_{LBB} from Theorem 2.5.2.

Theorem 3.2.17 (Alternative definition of C_{LBB})**.** *The constant $0 < C_{\mathrm{LBB}}$ equals*

$$C_{\mathrm{LBB}} = \inf_{z^\perp \in Z^\perp \setminus \{0\}} \|\operatorname{div} z^\perp\|_{L^2(\Omega)} / \|\nabla z^\perp\|_{L^2(\Omega)}. \tag{3.67}$$

Proof. Gauss's divergence theorem leads to

$$\int_\Omega \operatorname{div} v \,\mathrm{d}x = \int_{\partial\Omega} v \cdot \nu \,\mathrm{d}x = 0 \qquad \text{for all } v \in H^1_0(\Omega;\mathbb{R}^d).$$

Therefore, the linear operator $\operatorname{div} : H^1_0(\Omega;\mathbb{R}^d) \to L^2_0(\Omega) := \{q \in L^2(\Omega) \mid \int_\Omega q \,\mathrm{d}x = 0\}$. Theorem 2.5.2 shows that $\operatorname{div} : H^1_0(\Omega;\mathbb{R}^d) \to L^2_0(\Omega)$ is surjective. The kernel of $\operatorname{div} : H^1_0(\Omega;\mathbb{R}^d) \to L^2_0(\Omega)$ equals Z and so $\operatorname{div}|_{Z^\perp} : Z^\perp \to L^2_0(\Omega)$ is bijective. This results in the equivalence of (2.25) and

$$C_{\mathrm{LBB}} = \inf_{z^\perp \in Z^\perp \setminus \{0\}} \sup_{v \in H^1_0(\Omega;\mathbb{R}^d)\setminus\{0\}} \frac{(\operatorname{div} z^\perp, \operatorname{div} v)_{L^2(\Omega)}}{\|\operatorname{div} z^\perp\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)}} = \inf_{z^\perp \in Z^\perp \setminus \{0\}} \frac{\|\operatorname{div} z^\perp\|_{L^2(\Omega)}}{\|\nabla z^\perp\|_{L^2(\Omega)}}. \qquad \square$$

Let $\gamma > 0$ be an arbitrary weight and let $a(\bullet,\bullet)$ and $(\bullet,\bullet)_X$ induce the norms

$$\begin{aligned} \|(v,\tau)\|_a &:= LS(0;v,\tau)^{1/2} = (\|\operatorname{dev}\tau - \nabla v\|^2_{L^2(\Omega)} + \gamma \|\operatorname{div}\tau\|^2_{L^2(\Omega)})^{1/2}, \\ \|(v,\tau)\|_X &:= (\|\operatorname{dev}\tau\|^2_{L^2(\Omega)} + \|\nabla v\|^2_{L^2(\Omega)} + \gamma \|\operatorname{div}\tau\|^2_{L^2(\Omega)})^{1/2} \qquad \text{for all } (v,\tau) \in X. \end{aligned} \tag{3.68}$$

Lemma 3.2.18 (Orthogonal decomposition of X)**.** *Define the spaces $X_1 := Z \times (\Sigma_1 \oplus \Sigma_2)$ and $X_2 := Z^\perp \times \Sigma_0$ with $\Sigma_0, \Sigma_1, \Sigma_2$ from (3.59). The decomposition $X = X_1 \oplus X_2$ is orthogonal with respect to the inner products $a(\bullet,\bullet)$ and $(\bullet,\bullet)_X$, that is*

$$a(x_1,x_2) = 0 = (x_1,x_2)_X \qquad \textit{for all } x_1 \in X_1 \textit{ and } x_2 \in X_2.$$

Proof. Let $x_1 = (z,\tau) \in X_1$ and $x_2 = (z^\perp, \vartheta_0) \in X_2$. The definition of $Z^\perp$ and Theorem 3.2.10 imply $(x_1,x_2)_X = (\nabla z, \nabla z^\perp)_{L^2(\Omega)} + (\operatorname{dev}\tau, \operatorname{dev}\vartheta_0)_{L^2(\Omega)} + \gamma\,(\operatorname{div}\tau, \operatorname{div}\vartheta_0)_{L^2(\Omega)} = 0$. Similar arguments prove

$$\begin{aligned} a(x_1,x_2) &= (\operatorname{dev}\tau - \nabla z, \operatorname{dev}\vartheta_0 - \nabla z^\perp)_{L^2(\Omega)} + \gamma\,(\operatorname{div}\tau, \operatorname{div}\vartheta_0)_{L^2(\Omega)} \\ &= -(\nabla z, \operatorname{dev}\vartheta_0)_{L^2(\Omega)} - (\operatorname{dev}\tau, \nabla z^\perp)_{L^2(\Omega)}. \end{aligned} \tag{3.69}$$

The identity $\mathrm{tr}(\nabla z) = \operatorname{div} z = 0$ implies $\operatorname{dev}(\nabla z) = \nabla z$. This, an integration by parts, the property of the deviator (3.60a), and $\operatorname{div} \vartheta_0 = 0$ result in

$$(\nabla z, \operatorname{dev} \vartheta_0)_{L^2(\Omega)} = (\operatorname{dev}(\nabla z), \vartheta_0)_{L^2(\Omega)} = (\nabla z, \vartheta_0)_{L^2(\Omega)} = -(z, \operatorname{div} \vartheta_0)_{L^2(\Omega)} = 0. \tag{3.70}$$

The eigenfunctions ϕ_k from Theorem 3.2.10 satisfy $\operatorname{dev} \phi_k = -\lambda_k^{-1} \nabla \operatorname{div} \phi_k$ for all $k \in \mathbb{N}$. Lemma 3.2.14 shows $\operatorname{div} \phi_k \in Z$ and so $(\operatorname{dev} \phi_k, \nabla z^\perp)_{L^2(\Omega)} = -\lambda_k^{-1} (\nabla \operatorname{div} \phi_k, \nabla z^\perp)_{L^2(\Omega)} = 0$ for all $k \in \mathbb{N}$. This orthogonality, the split $\tau = \tau_1 + \tau_2$ with $\tau_1 \in \Sigma_1$, $\tau_2 \in \Sigma_2$, and the density of $\operatorname{span}\{\phi_k \mid k \in \mathbb{N}\}$ in Σ_2 yield

$$(\operatorname{dev} \tau, \nabla z^\perp)_{L^2(\Omega)} = (\operatorname{dev} \tau_2, \nabla z^\perp)_{L^2(\Omega)} = 0. \tag{3.71}$$

The combination of (3.69)–(3.71) proves $a(x_1, x_2) = 0$. □

Lemma 3.2.19 (Properties of $Z^\perp$)**.**
(i) Let $z^\perp \in Z^\perp$, then there exists a unique function $\xi_0 \in \Sigma_0$ with $\operatorname{dev} \xi_0 = \operatorname{dev} \nabla z^\perp$.
(ii) Let $z^\perp \in Z^\perp \setminus \{0\}$ with $C_{\mathrm{LBB}} = \|\operatorname{div} z^\perp\|_{L^2(\Omega)} / \|\nabla z^\perp\|_{L^2(\Omega)}$. Then $\xi_0 \in \Sigma_0$ with $\operatorname{dev} \xi_0 = \operatorname{dev} \nabla z^\perp \neq 0$ from (i) reads $\xi_0 = \nabla z^\perp - C_{\mathrm{LBB}}^{-2} \operatorname{div} z^\perp I_{d\times d}$ and

$$(\nabla z^\perp, \nabla v)_{L^2(\Omega)} = C_{\mathrm{LBB}}^{-2} (\operatorname{div} z^\perp, \operatorname{div} v)_{L^2(\Omega)} \qquad \text{for all } v \in H_0^1(\Omega; \mathbb{R}^d). \tag{3.72}$$

Proof. Step 1 (Proof of (i)). Let $z^\perp \in Z^\perp$. Define the unique Riesz representation ξ_0 in the Hilbert space Σ_0 with

$$(\operatorname{dev} \xi_0, \operatorname{dev} \tau_0)_{L^2(\Omega)} = (\operatorname{dev} \nabla z^\perp, \operatorname{dev} \tau_0)_{L^2(\Omega)} \qquad \text{for all } \tau_0 \in \Sigma_0. \tag{3.73}$$

The arguments in the end of the proof of Lemma 3.2.18 show $(\nabla z^\perp, \operatorname{dev} \tau)_{L^2(\Omega)} = 0$ for all $\tau \in \Sigma_1 \oplus \Sigma_2$. This, the orthogonality of Σ_0, Σ_1, and Σ_2 from Theorem 3.2.10, (3.60a), $\operatorname{dev} I_{d\times d} = 0$, and (3.73) result in

$$(\operatorname{dev} \xi_0, \tau)_{L^2(\Omega)} = (\operatorname{dev} \nabla z^\perp, \tau)_{L^2(\Omega)} \qquad \text{for all } \tau \in H(\operatorname{div}, \Omega; \mathbb{R}^{d\times d}). \tag{3.74}$$

Since $H(\operatorname{div}, \Omega; \mathbb{R}^{d\times d})$ is dense in $L^2(\Omega; \mathbb{R}^{d\times d})$, the identity in (3.74) yields

$$\operatorname{dev} \xi_0 = \operatorname{dev} \nabla z^\perp. \tag{3.75}$$

The uniqueness of the Riesz representation $\xi_0 \in \Sigma_0$ with (3.73) shows that any function $\vartheta_0 \in \Sigma_0$ with $\operatorname{dev} \vartheta_0 = \operatorname{dev} \nabla z^\perp$ equals ξ_0.

Step 2 ($\operatorname{dev} \nabla z^\perp \neq 0$). Suppose $z^\perp \in Z^\perp \setminus \{0\}$ satisfies $C_{\mathrm{LBB}} = \|\operatorname{div} z^\perp\|_{L^2(\Omega)} / \|\nabla z^\perp\|_{L^2(\Omega)}$. Assume $\operatorname{dev} \nabla z^\perp = 0$, then (3.60b) results in $\|\nabla z^\perp\|_{L^2(\Omega)}^2 = d^{-1} \|\operatorname{div} z^\perp\|_{L^2(\Omega)}^2$. The combination with $C_{\mathrm{LBB}} = \|\operatorname{div} z^\perp\|_{L^2(\Omega)} / \|\nabla z^\perp\|_{L^2(\Omega)}$ implies $1 < d = C_{\mathrm{LBB}}^2$. But (3.67) shows $C_{\mathrm{LBB}} \leq 1$. This contradiction proves $\operatorname{dev} \nabla z^\perp \neq 0$.

Step 3 (Proof of (3.72)). Suppose $z^\perp \in Z^\perp \setminus \{0\}$ satisfies $C_{\mathrm{LBB}} = \|\operatorname{div} z^\perp\|_{L^2(\Omega)} / \|\nabla z^\perp\|_{L^2(\Omega)}$. The Riesz representation theorem yields the existence of a function $v \in H_0^1(\Omega; \mathbb{R}^d)$ with

$$(\nabla v, \nabla w)_{L^2(\Omega)} = C_{\mathrm{LBB}}^{-2} (\operatorname{div} z^\perp, \operatorname{div} w)_{L^2(\Omega)} \qquad \text{for all } w \in H_0^1(\Omega; \mathbb{R}^d). \tag{3.76}$$

Since $\operatorname{div} w = 0$ for all $w \in Z$, (3.76) implies $(\nabla v, \nabla w)_{L^2(\Omega)} = 0$ for all $w \in Z$ and so $v \in Z^\perp$. Thus, (3.67) shows $C_{\mathrm{LBB}}\|\nabla v\|_{L^2(\Omega)} \leq \|\operatorname{div} v\|_{L^2(\Omega)}$. This inequality, the identity $C_{\mathrm{LBB}}\|\nabla z^\perp\|_{L^2(\Omega)} = \|\operatorname{div} z^\perp\|_{L^2(\Omega)}$, the Cauchy-Schwarz inequality, and (3.76) show

$$\begin{aligned}\|\nabla v\|^2_{L^2(\Omega)} &= C^{-2}_{\mathrm{LBB}}(\operatorname{div} z^\perp, \operatorname{div} v)_{L^2(\Omega)} \leq C^{-2}_{\mathrm{LBB}}\|\operatorname{div} z^\perp\|_{L^2(\Omega)}\|\operatorname{div} v\|_{L^2(\Omega)} \\ &\leq \|\nabla v\|_{L^2(\Omega)}\|\nabla z^\perp\|_{L^2(\Omega)}.\end{aligned} \tag{3.77}$$

Similar arguments and the inequality $\|\nabla v\|_{L^2(\Omega)} \leq \|\nabla z^\perp\|_{L^2(\Omega)}$ from (3.77) prove

$$\begin{aligned}\|\nabla z^\perp\|^2_{L^2(\Omega)} &= C^{-2}_{\mathrm{LBB}}\|\operatorname{div} z^\perp\|^2_{L^2(\Omega)} = (\nabla v, \nabla z^\perp)_{L^2(\Omega)} \\ &\leq \|\nabla v\|_{L^2(\Omega)}\|\nabla z^\perp\|_{L^2(\Omega)} \leq \|\nabla z^\perp\|^2_{L^2(\Omega)}.\end{aligned} \tag{3.78}$$

Thus, $\|\nabla z^\perp\|_{L^2(\Omega)} = \|\nabla v\|_{L^2(\Omega)}$ and, since the Cauchy-Schwarz inequality leads to the equality $(\nabla v, \nabla z^\perp)_{L^2(\Omega)} = \|\nabla v\|_{L^2(\Omega)}\|\nabla z^\perp\|_{L^2(\Omega)}$ if and only if the functions v and $z^\perp$ are linearly dependent, it holds $v = z^\perp$. Hence, $z^\perp$ satisfies (3.72).

Step 4 (Representation of ξ_0). Let $z^\perp \in Z^\perp \setminus \{0\}$ with $C_{\mathrm{LBB}} = \|\operatorname{div} z^\perp\|_{L^2(\Omega)}/\|\nabla z^\perp\|_{L^2(\Omega)}$ and define the function $\xi_0 := \nabla z^\perp - C^{-2}_{\mathrm{LBB}}\operatorname{div} z^\perp I_{d\times d}$. This definition shows $\operatorname{dev}\xi_0 = \operatorname{dev}\nabla z^\perp$. Equation (3.72) and the identity $\operatorname{tr}(\nabla v) = \operatorname{div} v$ result in

$$\begin{aligned}(\xi_0, \nabla v)_{L^2(\Omega)} &= (\nabla z^\perp, \nabla v)_{L^2(\Omega)} - C^{-2}_{\mathrm{LBB}}(\operatorname{div} z^\perp I_{d\times d}, \nabla v)_{L^2(\Omega)} \\ &= (\nabla z^\perp, \nabla v)_{L^2(\Omega)} - C^{-2}_{\mathrm{LBB}}(\operatorname{div} z^\perp, \operatorname{div} v)_{L^2(\Omega)} = 0 \qquad \text{for all } v \in H^1_0(\Omega, \mathbb{R}^d).\end{aligned}$$

Thus, the divergence $\operatorname{div}\xi_0 = 0$. The property $\operatorname{div} z^\perp \in L^2_0(\Omega)$ from the proof of Theorem 3.2.17 implies that the trace $\operatorname{tr}(\xi_0) = (1 - C^{-2}_{\mathrm{LBB}}d)\operatorname{div} z^\perp \in L^2_0(\Omega)$. The combination of these two properties implies $\xi_0 \in \Sigma_0$. The uniqueness of $\xi_0 \in \Sigma_0$ with $\operatorname{dev}\xi_0 = \operatorname{dev}\nabla z^\perp$ concludes the proof. □

The following main result of this section computes the equivalence constants of the norms $\|\bullet\|_a$ and $\|\bullet\|_X$ with given weight $\gamma > 0$ from (3.68). Recall the smallest eigenvalue λ_1 from Theorem 3.2.10, the LBB constant C_{LBB} from Theorem 2.5.2, and the dimension $2 \leq d \in \mathbb{N}$ with $\Omega \subset \mathbb{R}^d$.

Theorem 3.2.20 (Ellipticity constants α and β)**.** *It holds*

$$\alpha := \inf_{x\in X\setminus\{0\}} \|x\|^2_a/\|x\|^2_X = \min\{1 - (\gamma\lambda_1 + 1)^{-1/2}, 1 - (1 - C^2_{\mathrm{LBB}}/d)^{1/2}\}, \tag{3.79a}$$

$$\beta := \sup_{x\in X\setminus\{0\}} \|x\|^2_a/\|x\|^2_X = \max\{1 + (\gamma\lambda_1 + 1)^{-1/2}, 1 + (1 - C^2_{\mathrm{LBB}}/d)^{1/2}\}. \tag{3.79b}$$

Proof. This proof focuses on the computation of α. The identity in (3.79b) follows analogously. The orthogonal decomposition $X = X_1 \oplus X_2$ from Lemma 3.2.18 implies $\alpha = \min\{\alpha_1, \alpha_2\}$ with

$$\alpha_k := \inf_{x_k\in X_k\setminus\{0\}} \|x_k\|^2_a/\|x_k\|^2_X \qquad \text{for } k = 1, 2.$$

Step 1 (Computation of α_1). Since $X_1 \subset X_{\mathrm{red}}$ with X_{red} from (3.62), the combination of Theorem 3.2.15, Theorem 3.2.16, and Theorem 3.1.3 proves

$$\mu_1 := 1 - (\gamma\lambda_1 + 1)^{-1/2} = \inf_{x\in X_{\mathrm{red}}\setminus\{0\}} \|x\|^2_a/\|x\|^2_X \leq \inf_{x\in X_1\setminus\{0\}} \|x\|^2_a/\|x\|^2_X = \alpha_1.$$

This bound and the identity $\mu_1 = \|\psi_1\|_a^2/\|\psi_1\|_X^2$ with $\psi_1 \in X_1$ from (3.64) result in $\alpha_1 = \mu_1$.

Step 2 (Lower bound for α_2). Let $z^\perp \in Z^\perp\backslash\{0\}$. The inequality $\|\nabla z^\perp\|_{L^2(\Omega)}\|\operatorname{dev}\tau_0\|_{L^2(\Omega)} \leq (\|\nabla z^\perp\|_{L^2(\Omega)}^2 + \|\operatorname{dev}\tau_0\|_{L^2(\Omega)}^2)/2$ for all $\tau_0 \in \Sigma_0$ yields

$$\sup_{\tau_0\in\Sigma_0\backslash\{0\}} \frac{2(\nabla z^\perp, \operatorname{dev}\tau_0)_{L^2(\Omega)}}{\|\nabla z^\perp\|_{L^2(\Omega)}^2 + \|\operatorname{dev}\tau_0\|_{L^2(\Omega)}^2} \leq \sup_{\tau_0\in\Sigma_0\backslash\{0\}} \frac{(\nabla z^\perp, \operatorname{dev}\tau_0)_{L^2(\Omega)}}{\|\nabla z^\perp\|_{L^2(\Omega)}\,\|\operatorname{dev}\tau_0\|_{L^2(\Omega)}}. \tag{3.80}$$

Given $\xi_0 \in \Sigma_0\backslash\{0\}$, set the constant $\kappa := \|\nabla z^\perp\|_{L^2(\Omega)}/\|\operatorname{dev}\xi_0\|_{L^2(\Omega)}$. Since $\|\operatorname{dev}\kappa\xi_0\|_{L^2(\Omega)} = \|\nabla z^\perp\|_{L^2(\Omega)}$, it holds $2\|\nabla z^\perp\|_{L^2(\Omega)}\|\operatorname{dev}\kappa\xi_0\|_{L^2(\Omega)} = \|\nabla z^\perp\|_{L^2(\Omega)}^2 + \|\operatorname{dev}\kappa\xi_0\|_{L^2(\Omega)}^2$. Thus,

$$\begin{aligned}\frac{(\nabla z^\perp, \operatorname{dev}\xi_0)_{L^2(\Omega)}}{\|\nabla z^\perp\|_{L^2(\Omega)}\,\|\operatorname{dev}\xi_0\|_{L^2(\Omega)}} &= \frac{2(\nabla z^\perp, \operatorname{dev}\kappa\xi_0)_{L^2(\Omega)}}{\|\nabla z^\perp\|_{L^2(\Omega)}^2 + \|\operatorname{dev}\kappa\xi_0\|_{L^2(\Omega)}^2}\\ &\leq \sup_{\tau_0\in\Sigma_0\backslash\{0\}} \frac{2(\nabla z^\perp, \operatorname{dev}\tau_0)_{L^2(\Omega)}}{\|\nabla z^\perp\|_{L^2(\Omega)}^2 + \|\operatorname{dev}\tau_0\|_{L^2(\Omega)}^2}.\end{aligned}$$

This inequality holds for all $\xi_0 \in \Sigma_0 \setminus \{0\}$ and so

$$\sup_{\tau_0\in\Sigma_0\backslash\{0\}} \frac{(\nabla z^\perp, \operatorname{dev}\tau_0)_{L^2(\Omega)}}{\|\nabla z^\perp\|_{L^2(\Omega)}\,\|\operatorname{dev}\tau_0\|_{L^2(\Omega)}} \leq \sup_{\tau_0\in\Sigma_0\backslash\{0\}} \frac{2(\nabla z^\perp, \operatorname{dev}\tau_0)_{L^2(\Omega)}}{\|\nabla z^\perp\|_{L^2(\Omega)}^2 + \|\operatorname{dev}\tau_0\|_{L^2(\Omega)}^2}. \tag{3.81}$$

The combination of the inequalities in (3.80)–(3.81) implies the identity

$$\sup_{\tau_0\in\Sigma_0\backslash\{0\}} \frac{2(\nabla z^\perp, \operatorname{dev}\tau_0)_{L^2(\Omega)}}{\|\nabla z^\perp\|_{L^2(\Omega)}^2 + \|\operatorname{dev}\tau_0\|_{L^2(\Omega)}^2} = \sup_{\tau_0\in\Sigma_0\backslash\{0\}} \frac{(\nabla z^\perp, \operatorname{dev}\tau_0)_{L^2(\Omega)}}{\|\nabla z^\perp\|_{L^2(\Omega)}\,\|\operatorname{dev}\tau_0\|_{L^2(\Omega)}}. \tag{3.82}$$

Lemma 3.2.19*(i)* implies the existence of a function $\xi_0 \in \Sigma_0$ with $\operatorname{dev}\xi_0 = \operatorname{dev}\nabla z^\perp$. This function and the identity in (3.82) lead to

$$\begin{aligned}\inf_{\tau_0\in\Sigma_0} \frac{\|(z^\perp,\tau_0)\|_a^2}{\|(z^\perp,\tau_0)\|_X^2} &= 1 - \sup_{\tau_0\in\Sigma_0\backslash\{0\}} \frac{2(\nabla z^\perp, \operatorname{dev}\tau_0)_{L^2(\Omega)}}{\|\nabla z^\perp\|_{L^2(\Omega)}^2 + \|\operatorname{dev}\tau_0\|_{L^2(\Omega)}^2}\\ &= 1 - \sup_{\tau_0\in\Sigma_0\backslash\{0\}} \frac{(\operatorname{dev}\xi_0, \operatorname{dev}\tau_0)_{L^2(\Omega)}}{\|\nabla z^\perp\|_{L^2(\Omega)}\,\|\operatorname{dev}\tau_0\|_{L^2(\Omega)}} = 1 - \frac{\|\operatorname{dev}\nabla z^\perp\|_{L^2(\Omega)}}{\|\nabla z^\perp\|_{L^2(\Omega)}}.\end{aligned} \tag{3.83}$$

The equation in (3.60b), the identity $\operatorname{div} z^\perp = \operatorname{tr}(\nabla z^\perp)$, and Theorem 3.2.17 show

$$\begin{aligned}\|\nabla z^\perp\|_{L^2(\Omega)}^2 &= \|\operatorname{dev}\nabla z^\perp\|_{L^2(\Omega)}^2 + d^{-1}\|\operatorname{div} z^\perp\|_{L^2(\Omega)}^2\\ &\geq \|\operatorname{dev}\nabla z^\perp\|_{L^2(\Omega)}^2 + d^{-1}C_{\mathrm{LBB}}^2\|\nabla z^\perp\|_{L^2(\Omega)}^2.\end{aligned} \tag{3.84}$$

Thus, $\|\operatorname{dev}\nabla z^\perp\|_{L^2(\Omega)}^2/\|\nabla z^\perp\|_{L^2(\Omega)}^2 \leq 1 - C_{\mathrm{LBB}}^2/d$. The combination with (3.83) implies

$$1 - (1 - C_{\mathrm{LBB}}^2/d)^{1/2} \leq \alpha_2.$$

Step 3 (Upper bound for α_2). Let $(z_n^\perp)_{n\in\mathbb{N}} \subset Z^\perp \setminus \{0\}$ be a sequence with $\|\operatorname{div} z_n^\perp\|_{L^2(\Omega)} \searrow C_{\mathrm{LBB}}\|\nabla z_n^\perp\|_{L^2(\Omega)}$ as $n \to \infty$. Equation (3.84) shows $\|\operatorname{dev}\nabla z_n^\perp\|_{L^2(\Omega)}^2/\|\nabla z_n^\perp\|_{L^2(\Omega)}^2 \nearrow 1 - C_{\mathrm{LBB}}^2/d$ as $n \to \infty$. This and (3.83) imply

$$\alpha_2 \leq \inf_{\tau_0\in\Sigma_0} \frac{\|(z_n^\perp,\tau_0)\|_a^2}{\|(z_n^\perp,\tau_0)\|_X^2} = 1 - \frac{\|\operatorname{dev}\nabla z_n^\perp\|_{L^2(\Omega)}}{\|\nabla z_n^\perp\|_{L^2(\Omega)}} \searrow 1 - (1 - C_{\mathrm{LBB}}^2/d)^{1/2} \quad \text{as } n \to \infty. \quad \square$$

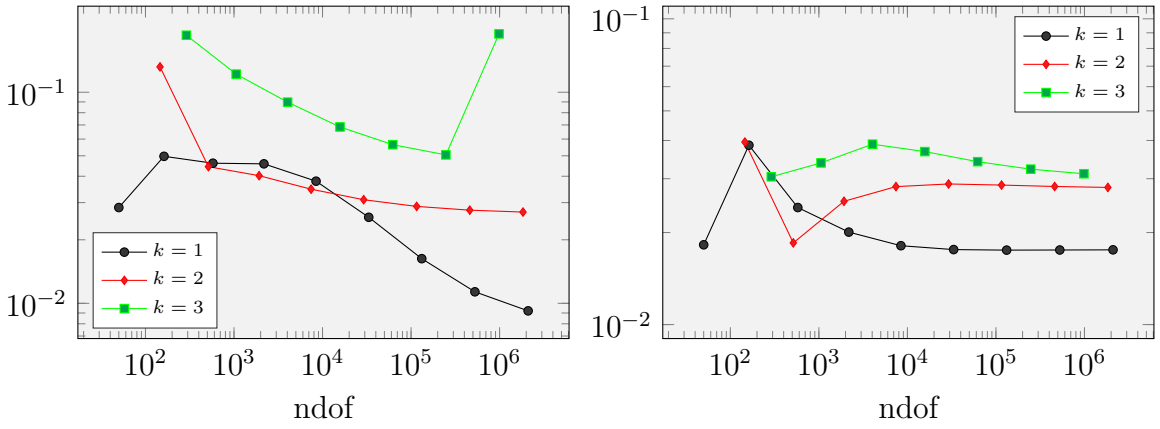


Figure 3.8: Distance $1 - LS(f;\mathfrak{u}_h)/\|\mathfrak{u}-\mathfrak{u}_h\|_X^2$ with weights $w_1 = 0$, $w_2 = 1$ (left) and $w_1 = 1$, $w_2 = 0$ (right) for polynomial degrees $k = 1, 2, 3$

Numerical experiment

This section concludes with a numerical investigation of the ratio $LS(f;\mathfrak{u}_h)/\|\mathfrak{u}-\mathfrak{u}_h\|_X^2$ with the weighted least-squares functional from (3.57) and weight $\gamma = 1$, discrete solution $\mathfrak{u}_h = \arg\min_{x_h \in X_h} LS(f;x_h)$, and exact solution $\mathfrak{u} = \arg\min_{x \in X} LS(f;x)$ to the Stokes problem with given right-hand side $f \in L^2(\Omega;\mathbb{R}^d)$. The domain $\Omega = (0,1)^2$ is the unit square. Define the functions [Lin07, Sec. 9.1]

$$v(x,y) := \begin{pmatrix} 2x^2(x-1)^2y(y-1)(2y-1) \\ -2y^2(y-1)^2x(x-1)(2x-1) \end{pmatrix} \in Z, \quad p(x,y) := (x^3+y^3-0.5) \in L^2_0(\Omega).$$

Let $w_1, w_2 \in \mathbb{R}$ be weights and set the right-hand side $f := -w_1 \operatorname{div} \nabla v + w_2 \nabla p \in L^2(\Omega;\mathbb{R}^d)$, then (2.21) shows that $\mathfrak{u} = (w_1 v, \sigma) \in X$ with $\sigma := w_1 \nabla v - w_2 p I_{2\times 2}$ solves the Stokes problem (2.22). Given a regular triangulation $\mathcal{T}$ and $k \in \mathbb{N}$, recall the Courant and Raviart-Thomas finite element spaces $S^k_0(\mathcal{T})$ and $RT_{k-1}(\mathcal{T})$ from (3.40) and define

$$\begin{aligned} S^k_0(\mathcal{T};\mathbb{R}^2) &:= \{v_h \in H^1_0(\Omega;\mathbb{R}^2) \mid v_h = (v_{h,1}, v_{h,2}) \text{ with } v_{h,1}, v_{h,2} \in S^k_0(\mathcal{T})\}, \\ RT_{k-1}(\mathcal{T};\mathbb{R}^{2\times 2}) &:= \{q_h \in H(\operatorname{div},\Omega;\mathbb{R}^{2\times 2}) \mid q_h = (q_{h,1}, q_{h,2}) \text{ with } q_{h,1}, q_{h,2} \in RT_{k-1}(\mathcal{T})\}. \end{aligned} \tag{3.85}$$

Figure 3.8 displays the term $1-LS(f;\mathfrak{u}_h)/\|\mathfrak{u}-\mathfrak{u}_h\|_X^2$ with discrete space $X_h := S^1_0(\mathcal{T};\mathbb{R}^2) \times (RT_{k-1}(\mathcal{T};\mathbb{R}^{2\times 2}) \cap \Sigma) \subset X = H^1_0(\Omega;\mathbb{R}^2) \times \Sigma$ (but $X_h \not\subset X_{\text{red}} = Z \times \Sigma$), uniformly refined triangulations $\mathcal{T}$, polynomial degrees $k = 1, 2, 3$, and different weights w_1, w_2. In all computations the term $1-LS(f;\mathfrak{u}_h)/\|\mathfrak{u}-\mathfrak{u}_h\|_X^2$ is positive, that is, the ratio $LS(f;\mathfrak{u}_h)/\|\mathfrak{u}-\mathfrak{u}_h\|_X^2 < 1$. The convergence history plot on the left-hand side of Figure 3.8 suggests the convergence $1-LS(f;\mathfrak{u}_h)/\|\mathfrak{u}-\mathfrak{u}_h\|_X^2 \to 0$ for $k = 1$. The ratio $LS(f;\mathfrak{u}_h)/\|\mathfrak{u}-\mathfrak{u}_h\|_X^2$ seems to converge towards 0.974 for $k = 2$. The convergence history plot on the right-hand side of Figure 3.8 indicates the convergence of the ratio $LS(f;\mathfrak{u}_h)/\|\mathfrak{u}-\mathfrak{u}_h\|_X^2$ towards values smaller than one as well. This suggest that the ratio $LS(f;\mathfrak{u}_h)/\|\mathfrak{u}-\mathfrak{u}_h\|_X^2$ does not tend to one in general. In other words, the asymptotic exactness of the least-squares residual does in general not apply to the LSFEM of this section with discrete ansatz spaces $X_h \not\subset X_{\text{red}}$.

4 Computation of the LBB constant

The Ladyzhenskaya-Babuška-Brezzi (LBB) constant $C_{\mathrm{LBB}} > 0$ from (2.25) is a key in the existence and stability of solutions in fluid dynamics. Its value enters reliability constants (as for example the combination of Theorem 3.2.20 and (3.2) shows) and influences the convergence rate of iterative algorithms (as for example pointed out in [Cro97, Eq. 0.3] for the Uzsawa's algorithm). The squared LBB constant equals the absolute value of the smallest non-zero element in the spectrum of the (non-compact) Cosserat operator [Vel96]. The abstract of [BCDG16] points out that "this eigenvalue problem does not fall into a class for which standard results about numerical approximations can be applied. Indeed, many reasonable finite element methods do not yield a convergent approximation".

4.1 A convergent scheme

Theorem 3.2.20 offers a simple ansatz to circumvent the difficulties in the approximation of the LBB constant. More precisely, the relation of the LBB constant and the coercivity constant from Theorem 3.2.20 allows to design a numerical scheme which results in a convergent approximation of the LBB constant. The numerical scheme bases on the computation of the following discrete inf-sup constant α_h. Recall the space X from (3.56) with the inner products $a(\bullet,\bullet)$ and $(\bullet,\bullet)_X$ and induced norms $\|\bullet\|_a$ and $\|\bullet\|_X$ from (3.68) with weight $\gamma > 0$. Let $X_h \subset X$ be a (discrete) subspace and define the coercivity constants

$$\alpha := \inf_{x\in X\setminus\{0\}} \frac{a(x,x)}{(x,x)_X} \leq \alpha_h := \inf_{x_h\in X_h\setminus\{0\}} \frac{a(x_h,x_h)}{(x_h,x_h)_X}. \tag{4.1}$$

Let λ_1 be the smallest eigenvalue from Theorem 3.2.10 and let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain with $2 \leq d \in \mathbb{N}$. Assume

$$(1+\gamma\lambda_1)^{-1} \leq 1 - C_{\mathrm{LBB}}^2/d. \tag{4.2}$$

Theorem 4.1.1 (Approximation of C_{LBB})**.** *Suppose (4.2).*
(i) If the discrete subspace X_h satisfies the density property (**D**) *on page 16, it holds*

$$C_{\mathrm{LBB}}^2 = d(1-(1-\alpha)^2) \leq C_{\mathrm{LBB},h}^2 := d(1-(1-\alpha_h)^2) \searrow C_{\mathrm{LBB}}^2 \qquad \text{as } h \to 0.$$

(ii) If there exists a function $z^\perp \in Z^\perp \setminus \{0\}$ with $C_{\mathrm{LBB}}\|\nabla z^\perp\|_{L^2(\Omega)} = \|\operatorname{div} z^\perp\|_{L^2(\Omega)}$, set $\xi_0 \in \Sigma_0$ as in Lemma 3.2.19(ii) and $\kappa := \|\nabla z^\perp\|_{L^2(\Omega)}/\|\operatorname{dev}\xi_0\|_{L^2(\Omega)}$. The normed function

$$x_{\min} := \|(z^\perp,\kappa\xi_0)\|_X^{-1}(z^\perp,\kappa\xi_0) \in X \tag{4.3}$$

minimizes (4.1) in the sense that $\|x_{\min}\|_X^2 = 1$ and $\|x_{\min}\|_a^2 = \alpha$. Moreover, it holds

$$0 \leq C_{\mathrm{LBB},h}^2 - C_{\mathrm{LBB}}^2 \leq 4d(1-d^{-1}C_{\mathrm{LBB}}^2)^{1/2} \min_{x_h\in X_h} \|x_{\min}-x_h\|_X^2. \tag{4.4}$$

Proof. Step 1 (Proof of (i)). Let $(x_n)_{n\in\mathbb{N}} \subset X$ be a sequence with $\|x_n\|_X = 1$ for all $n \in \mathbb{N}$ and $\|x_n\|_a^2 \to \alpha$ as $n \to \infty$. The density property (**D**) and the equivalence of $\|\bullet\|_X$ and $\|\bullet\|_a$ imply the existence of positive parameters $H_1 \geq H_2 \geq \ldots$ such that for all $n \in \mathbb{N}$ and $h \leq H_n$ exists a function $x_h \in X_h$ with $\|x_h\|_X^2 - 1/n \leq \|x_n\|_X^2$ and $\|x_n\|_a^2 \leq \|x_h\|_a^2 + 1/n$. Hence

$$\alpha \leq \lim_{h\to 0} \inf_{x_h \in X_h\setminus\{0\}} \frac{a(x_h,x_h)}{(x_h,x_h)_X} \leq \lim_{n\to\infty} \frac{\|x_n\|_a^2 + 1/n}{\|x_n\|_X^2 - 1/n} = \lim_{n\to\infty} \|x_n\|_a^2 = \alpha.$$

This proves $\alpha_h \searrow \alpha$ as $h \to 0$. Theorem 3.2.20 and (4.2) imply $\alpha = 1 - (1 - C_{\mathrm{LBB}}^2/d)^{1/2} \leq \alpha_h \leq \beta = 1 + (1 - C_{\mathrm{LBB}}^2/d)^{1/2}$. Thus, $C_{\mathrm{LBB}}^2 = d(1-(1-\alpha)^2) \leq C_{\mathrm{LBB},h}^2 := d(1-(1-\alpha_h)^2)$. Moreover, the convergence $\alpha_h \searrow \alpha$ results in $C_{\mathrm{LBB},h}^2 \searrow C_{\mathrm{LBB}}^2$ as $h \to 0$.

Step 2 (Minimizer $x_{\min}$). If there exists a function $z^\perp \in Z^\perp \setminus \{0\}$ with $\|\operatorname{div} z^\perp\|_{L^2(\Omega)} = C_{\mathrm{LBB}}\|\nabla z^\perp\|_{L^2(\Omega)}$, Lemma 3.2.19*(ii)* implies the existence of a function $\xi_0 \in \Sigma_0$ with $0 \neq \operatorname{dev}\xi_0 = \operatorname{dev}\nabla z^\perp$. Set $\kappa := \|\nabla z^\perp\|_{L^2(\Omega)}/\|\operatorname{dev}\xi_0\|_{L^2(\Omega)}$. The properties in (3.60), $\operatorname{tr}(\nabla z^\perp) = \operatorname{div} z^\perp$, $\operatorname{dev}\xi_0 = \operatorname{dev}\nabla z^\perp$, $\|\nabla z^\perp\|_{L^2(\Omega)} = \|\kappa \operatorname{dev}\xi_0\|_{L^2(\Omega)}$, and (4.2) prove

$$\begin{aligned}
\frac{\|(z^\perp,\kappa\xi_0)\|_a^2}{\|(z^\perp,\kappa\xi_0)\|_X^2} &= 1 - \frac{2\kappa(\nabla z^\perp, \operatorname{dev}\xi_0)_{L^2(\Omega)}}{\|\nabla z^\perp\|_{L^2(\Omega)}^2 + \|\kappa\operatorname{dev}\xi_0\|_{L^2(\Omega)}^2} = 1 - \frac{\kappa\|\operatorname{dev}\nabla z^\perp\|_{L^2(\Omega)}^2}{\|\nabla z^\perp\|_{L^2(\Omega)}^2} \\
&= 1 - \frac{\|\operatorname{dev}\nabla z^\perp\|_{L^2(\Omega)}}{\|\nabla z^\perp\|_{L^2(\Omega)}} = 1 - \frac{(\|\nabla z^\perp\|_{L^2(\Omega)}^2 - d^{-1}\|\operatorname{div} z^\perp\|_{L^2(\Omega)}^2)^{1/2}}{\|\nabla z^\perp\|_{L^2(\Omega)}} \\
&= 1 - (1 - d^{-1}C_{\mathrm{LBB}}^2)^{1/2} = \alpha.
\end{aligned}$$

Thus, the function $x_{\min} := \|(z^\perp,\kappa\xi_0)\|_X^{-1}(z^\perp,\kappa\xi_0)$ satisfies $\|x_{\min}\|_X^2 = 1$ and $\|x_{\min}\|_a^2 = \alpha$. Define the Riesz representation ϑ with $a(\vartheta, x) = \alpha\,(x_{\min}, x)_X$ for all $x \in X$. The definition of ϑ, the inequality $\alpha\|x\|_X^2 \leq \|x\|_a^2$ for all $x \in X$, and the Cauchy-Schwarz inequality imply $\|\vartheta\|_a^2 = \alpha\,(x_{\min},\vartheta)_X \leq \alpha\|x_{\min}\|_X\|\vartheta\|_X \leq \|x_{\min}\|_a\|\vartheta\|_a$ and so $\|\vartheta\|_a \leq \|x_{\min}\|_a$. Similar arguments and $\|\vartheta\|_a \leq \|x_{\min}\|_a$ show

$$\|x_{\min}\|_a^2 = \alpha\|x_{\min}\|_X^2 = a(\vartheta, x_{\min}) \leq \|x_{\min}\|_a\|\vartheta\|_a \leq \|x_{\min}\|_a^2.$$

Therefore, $\|\vartheta\|_a = \|x_{\min}\|_a$ and, since the Cauchy-Schwarz inequality leads to the equality $a(x_{\min},\vartheta) = \|x_{\min}\|_a\|\vartheta\|_a$ if and only if the functions are linearly dependent, it holds $\vartheta = x_{\min}$. In other words, $x_{\min} \in X$ solves the eigenvalue problem

$$a(x_{\min}, x) = \alpha\,(x_{\min}, x)_X \qquad \text{for all } x \in X. \tag{4.5}$$

Step 3 (A priori estimate). Let $x_h \in X_h$ be the best approximation of the minimizer $x_{\min} := \|(z^\perp,\kappa\xi_0)\|_X^{-1}(z^\perp,\kappa\xi_0) \in X$ from Step 2 in the sense that $\|x_{\min} - x_h\|_X = \min_{y_h\in X_h}\|x_{\min} - y_h\|_X$. The characterization of best approximations from Lemma 3.1.5 shows that this is equivalent to $(x_{\min} - x_h, y_h)_X = 0$ for all $y_h \in X_h$. This, (4.5), the Cauchy-Schwarz inequality, the equivalence of norms (3.79), and $\alpha = 1 - (1 - C_{\mathrm{LBB}}^2/d)^{1/2}$ and $\beta = 1 + (1 - C_{\mathrm{LBB}}^2/d)^{1/2}$ from Theorem 3.2.20 and (4.2) result with absolute value $|\bullet|$ in

$$\begin{aligned}
&|\|x_{\min}\|_a^2 - \|x_h\|_a^2| = |a(x_{\min}, x_{\min} - x_h) + a(x_h, x_{\min} - x_h)| \\
&= |2a(x_{\min}, x_{\min} - x_h) - \|x_{\min} - x_h\|_a^2| = |2\alpha\|x_{\min} - x_h\|_X^2 - \|x_{\min} - x_h\|_a^2| \\
&\leq \max\{\alpha, \beta - 2\alpha\}\|x_{\min} - x_h\|_X^2 = \alpha\|x_{\min} - x_h\|_X^2.
\end{aligned} \tag{4.6}$$

If $x_h \neq 0$, the inequality in (4.6), the triangle inequality, the Pythagorean theorem $\|x_{\min} - x_h\|_X^2 = 1 - \|x_h\|_X^2$, $\|x_h\|_a^2/\|x_h\|_X^2 \leq \beta$, and $\alpha + \beta = 2$ prove

$$\begin{aligned}\alpha_h - \alpha &\leq \frac{\|x_h\|_a^2}{\|x_h\|_X^2} - \|x_h\|_a^2 + \|x_h\|_a^2 - \|x_{\min}\|_a^2 \leq (1 - \|x_h\|_X^2)\frac{\|x_h\|_a^2}{\|x_h\|_X^2} + |\|x_h\|_a^2 - \|x_{\min}\|_a^2| \\ &\leq (\alpha + \beta)\|x_{\min} - x_h\|_X^2 = 2\|x_{\min} - x_h\|_X^2. \end{aligned} \tag{4.7}$$

If $x_h = 0$, $\alpha_h - \alpha \leq \beta - \alpha \leq 2$ and $\|x_{\min}\|_X = 1$ imply (4.7). The inequality in (4.7) and $\alpha = 1 - (1 - d^{-1}C_{\mathrm{LBB}}^2)^{1/2} \leq \alpha_h$ result in

$$\begin{aligned} C_{\mathrm{LBB},h}^2 - C_{\mathrm{LBB}}^2 &= d(1 - (1 - \alpha_h)^2) - d(1 - (1 - \alpha)^2) = d(2(\alpha_h - \alpha) + (\alpha^2 - \alpha_h^2)) \\ &= d(2 - \alpha - \alpha_h)(\alpha_h - \alpha) \leq 4d(1 - \alpha)\|x_{\min} - x_h\|_X^2 \\ &= 4d(1 - d^{-1}C_{\mathrm{LBB}}^2)^{1/2}\|x_{\min} - x_h\|_X^2. \end{aligned}$$

□

Theorem 4.1.1 proves the convergence of the numerical scheme for discrete spaces X_h with the density property (**D**). To investigate the rate of convergence, assume that there exist for all $s > 0$ and all $x \in X \cap (H^{1+s}(\Omega;\mathbb{R}^d) \times H^s(\Omega;\mathbb{R}^{d\times d}))$ some h-independent constant $C(x,s) > 0$ with

$$\min_{x_h \in X_h} \|x - x_h\|_X \leq C(x,s)h^s \qquad \text{for all } h > 0. \tag{4.8}$$

Estimate (4.8) is well-established for standard discretizations X_h like discretizations with Courant and Raviart-Thomas elements [Bra07, Chap. 3.5], [Bar15, Lem. 3.6] (where the parameter $h > 0$ refers to the maximal mesh-size of the underlying triangulation).

Theorem 4.1.2 (Rate of convergence)**.** *Let $s > 0$ and $z^\perp \in (Z^\perp \cap H^{1+s}(\Omega;\mathbb{R}^d)) \setminus \{0\}$ with $C_{\mathrm{LBB}}\|\nabla z^\perp\|_{L^2(\Omega)} = \|\operatorname{div} z^\perp\|_{L^2(\Omega)}$. Suppose (4.2) and (4.8). Then there exists an h-independent constant $C < \infty$ with*

$$0 \leq C_{\mathrm{LBB},h}^2 - C_{\mathrm{LBB}}^2 \leq Ch^{2s} \qquad \textit{for all } h > 0.$$

Proof. The assumptions of this theorem imply that the minimizer from (4.3) satisfies $x_{\min} \in X \cap (H^{1+s}(\Omega;\mathbb{R}^d) \times H^s(\Omega;\mathbb{R}^{d\times d}))$. Thus, the application of (4.8) to (4.4) proves this theorem with $C = 4d(1 - d^{-1}C_{\mathrm{LBB}}^2)^{1/2}C(x_{\min},s)^2 < \infty$. □

Theorem 4.1.1–4.1.2 show that the computation of the discrete inf-sup constant allows for the approximation of the LBB constant. The rate of convergence is similar to the rate from [Gal19, Thm. 11]. A downside of the proposed method is that the validation of the assumption in (4.2) requires some a priori knowledge of λ_1 and C_{LBB}. The following theorem circumvents this downside by a computable a posteriori criterion which implies (4.2). The criterion involves the d-dimensional Lebesgue measure $|\Omega|$ and $|B_1(0)|$ of the domain $\Omega \subset \mathbb{R}^d$ and the unit ball $B_1(0) := \{x \in \mathbb{R}^d \mid \|x\|_2 < 1\}$ with Euclidean distance $\|\bullet\|_2$. Define the constant

$$C(\Omega) := 4\pi^2 d\,(2+d)^{-1}|\Omega|^{-2/d}|B_1(0)|^{-2/d}.$$

Theorem 4.1.3 (A posteriori criterion for (4.2))**.** *If*

$$\alpha_h := \inf_{x_h \in X_h \setminus \{0\}} \|x_h\|_a^2/\|x_h\|_X^2 \leq 1 - (\gamma C(\Omega) + 1)^{-1/2}, \tag{4.9}$$

then the weight $\gamma > 0$ satisfies the assumption in (4.2).

Proof. The constant $C(\Omega)$ is a lower bound for principal eigenvalue λ_1 from Theorem 3.2.10 [Ily09, Cor. 2.2.], that is

$$4\pi^2 d\,(2+d)^{-1}|\Omega|^{-2/d}|B_1(0)|^{-2/d} = C(\Omega) < \lambda_1.$$

This estimate and (4.9) imply

$$\alpha \leq \alpha_h \leq 1 - (\gamma C(\Omega) + 1)^{-1/2} < 1 - (\gamma\lambda_1 + 1)^{-1/2}. \tag{4.10}$$

The combination of (4.10) and $\alpha = \min\{1 - (\gamma\lambda_1 + 1)^{-1/2}, 1 - (1 - C_{\mathrm{LBB}}^2/d)^{1/2}\}$ from Theorem 3.2.20 shows $\alpha = 1 - (1 - C_{\mathrm{LBB}}^2/d)^{1/2}$ and so validates (4.2). □

Remark 4.1.4 (Choice of the weight γ)**.** *Let $x_h = (v_h, \tau_h) \in X_h \setminus \{0\}$ with $\|x_h\|_a^2/\|x_h\|_X^2 < 1$ for some weight γ, then the ratio*

$$\frac{\|x_h\|_a^2}{\|x_h\|_X^2} = \frac{\|\operatorname{dev}\tau_h - \nabla v_h\|_{L^2(\Omega)}^2 + \gamma\|\operatorname{div}\tau_h\|_{L^2(\Omega)}^2}{\|\nabla v_h\|_{L^2(\Omega)}^2 + \|\operatorname{dev}\tau_h\|_{L^2(\Omega)}^2 + \gamma\|\operatorname{div}\tau_h\|_{L^2(\Omega)}^2} < 1$$

increases monotonically in $\gamma > 0$. Thus, the choice of a smaller weight γ in (3.68) results in a smaller ellipticity constant α_h and so in a better approximation of C_{LBB} (under the assumption that γ still satisfies (4.2)).

Remark 4.1.5 (Guaranteed upper bounds for λ_j)**.** *If an upper bound $C_{\mathrm{LBB}}^{\mathrm{up}}$ of C_{LBB} is known, all eigenvalues $\mu_{h,j}$ from (4.11) with $\mu_{h,j} \leq 1 - (1 - (C_{\mathrm{LBB}}^{\mathrm{up}})^2/d)^{1/2} < 1$ allow for guaranteed and convergent upper bounds $\lambda_{h,j} := ((\mu_{h,j} - 1)^{-2} - 1)/\gamma$ of the eigenvalues λ_j from Theorem 3.2.10. In other words, $\lambda_{h,j} \searrow \lambda_j$ for all $j = 1, \ldots, \dim X_h$ with $\mu_{h,j} \leq 1 - (1 - (C_{\mathrm{LBB}}^{\mathrm{up}})^2/d)^{1/2} < 1$. Established methods approximate λ_j by solving an eigenvalue problem in mixed form (see for example [MORR81]). In contrast to the proposed method, these approaches do in general not allow for guaranteed upper eigenvalue bounds and the matrices in the discrete eigenvalue problem are not positive definite.*

Remark 4.1.6 (Existing eigenvalue approximations with LSFEMs)**.** *Bramble, Kolev, and Pasciak design a least-squares problem with eigenvalues that equal the Maxwell eigenvalues from Theorem 2.4.2. This least-squares problem allows for convergent eigenvalue approximations [BKP05a]. A difference to the approach of this thesis (which allows for the approximation of the Maxwell eigenvalue with the arguments of this section and the results from Section 3.2.2 as well) is that this thesis utilizes existing LSFEMs and investigates the relation of the eigenvalues in the LSFEM and the eigenvalues of the underlying problem.*

Barrenechea, Boulton, and Boussaïd introduce a further related approach in [BBB14, BBB17]. They shift the Maxwell eigenvalues and utilize a squared (energy) functional to compute upper and lower eigenvalue bounds with an indefinite discrete eigenvalue problem.

4.2 Numerical experiments

The approximation of the LBB constant C_{LBB} with Theorem 4.1.1 allows to utilize the FEniCS code from the experiment in Section 3.2.3. More precisely, given the discrete space $X_h := S_0^k(\mathcal{T};\mathbb{R}^2) \times (RT_{k-1}(\mathcal{T};\mathbb{R}^{2\times 2}) \cap \Sigma) \subset X$ from (3.85) with regular triangulation $\mathcal{T}$ of the domain $\Omega \subset \mathbb{R}^2$ and polynomial degree $k \in \mathbb{N}$, the combination of the

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$
$\nu_{h,j}^{\mathrm{CCDL}}$	0.008129	0.031410	0.066825	0.110173	0.156691	0.199097
$\nu_{h,j}$	0.008129	0.031410	0.066825	0.110172	0.156638	0.191819

Table 4.1: Approximations of the first six Cosserat eigenvalues in Experiment 1

symmetric positive definite system matrices from the LSFEM solver and the algorithm in Appendix A.1.2 computes eigenvalues $0 < \mu_{h,1} \leq \mu_{h,2} \leq \cdots \leq \mu_{h,J}$ and eigenfunctions $\psi_{h,1}, \psi_{h,2}, \ldots, \psi_{h,J} \in X_h \setminus \{0\}$ with $J = \dim X_h$ and

$$a(\psi_{h,j}, x_h) = \mu_{h,j} \, (\psi_{h,j}, x_h)_X \qquad \text{for all } x_h \in X_h \text{ and } j = 1, \ldots, J. \tag{4.11}$$

The smallest discrete eigenvalue $\mu_{h,1} = \alpha_h$ results in the approximation $C^2_{\mathrm{LBB},h} := 2(1 - (1-\alpha_h)^2)$ with $C^2_{\mathrm{LBB},h} \searrow C^2_{\mathrm{LBB}}$ as the maximal mesh-size of $\mathcal{T}$ vanishes. The only further algorithm that allows for the systematic construction of monotone sequences that converge to the LBB constant has recently been introduced in [Gal19]. All experiments utilize the weight $\gamma = 1$ and satisfy the a posteriori criterion (4.9).

Remark 4.2.1 (Inexact computation of $\mu_{h,1}, \mu_{h,2}, \ldots, \mu_{h,J}$)**.** *Any upper eigenvalue bound $\mu^{\mathrm{up}}_{h,1} \geq \mu_{h,1}$ with $\mu^{\mathrm{up}}_{h,1} \leq \beta_h = \sup_{x_h \in X_h \setminus \{0\}} \|x_h\|_a^2 / \|x_h\|_X^2$ results in an upper bound $d(1 - (1-\mu^{\mathrm{up}}_{h,1})^2) \geq d(1-(1-\alpha_h)^2) = C^2_{\mathrm{LBB},h} \geq C^2_{\mathrm{LBB}}$. Since the algorithm in Appendix A.1.2 computes upper bounds $\mu^{\mathrm{up}}_{h,j} \geq \mu_{h,j}$ with $\mu^{\mathrm{up}}_{h,j} \leq \beta$ for all $j = 1, \ldots, J$, the numerical experiments of this section ignore the fact that the approximation of the eigenvalues is inexact.*

Experiment 1 (Isolated eigenvalues). The numerical experiment in [CCDL15, Fig. 7] indicates the existence of six isolated eigenvalues $0 < \nu_1 = C^2_{\mathrm{LBB}} \leq \nu_2 \leq \cdots \leq \nu_6$ of the Cosserat operator for the rectangular domain $\Omega = (0,1) \times (0,10)$. The first experiment approximates these eigenvalues. It computes the eigenvalues $\mu_{h,1}, \ldots, \mu_{h,6}$ and eigenfunctions $\psi_{h,1}, \ldots, \psi_{h,6} \in X_h := S^3_0(\mathcal{T}; \mathbb{R}^2) \times (RT_2(\mathcal{T}; \mathbb{R}^{2\times 2}) \cap \Sigma)$ with (4.11). The adaptively refined triangulations $\mathcal{T}$ result from Algorithm 3 on page 137 for all $j = 1, \ldots, 6$. More precisely, given $j = 1, \ldots, 6$, the eigenfunction $\psi_{h,j} = (v_h, \xi_h) \in X_h$, and the weight $\kappa = \|\nabla v_h\|_{L^2(\Omega)} / \|\operatorname{dev} \nabla v_h\|_{L^2(\Omega)}$, the adaptive mesh refinement applies Algorithm 3 with bulk parameter $\Theta = 0.3$ and refinement indicator (motivated by the properties of the minimizer $x_{\min} \in X$ from (4.3))

$$\eta^2(T) := \kappa^{-2} \|\operatorname{div} \xi_h\|^2_{L^2(T)} + \|\operatorname{dev}(\nabla v_h - \kappa^{-1} \xi_h)\|^2_{L^2(T)} \qquad \text{for all } T \in \mathcal{T}. \tag{4.12}$$

The algorithm stops after $\mathrm{ndof} = \dim X_h$ exceeds 10^5. Table 4.1 displays the approximations $\nu^{\mathrm{CCDL}}_{h,j}$ from [CCDL15, Fig. 7] and $\nu_{h,j} := 2(1-(1-\mu_{h,j})^2)$ on the finest triangulation $\mathcal{T}$ for $j = 1, \ldots, 6$. The approximations $\nu_{h,j}$ decreases monotonically as ndof increases. This suggests convergence from above, that is $\nu_{h,j} \searrow \nu_j$ for all $j = 1, \ldots, 6$. This observation extends the theoretical result from Theorem 4.1.1, which states solely the convergence of the first eigenvalue $\nu_{h,1} \searrow \nu_1 = C^2_{\mathrm{LBB}}$ as the mesh-size $h \to 0$. If $\nu_j \leq \nu_{h,j}$ for $j = 1, \ldots, 6$, the results improve the estimates in [CCDL15, Fig. 7] for $j = 4, 5, 6$.

Experiment 2 (Isolated eigenvalue). Let the rectangular domain $\Omega = (0,1) \times (0,2)$.

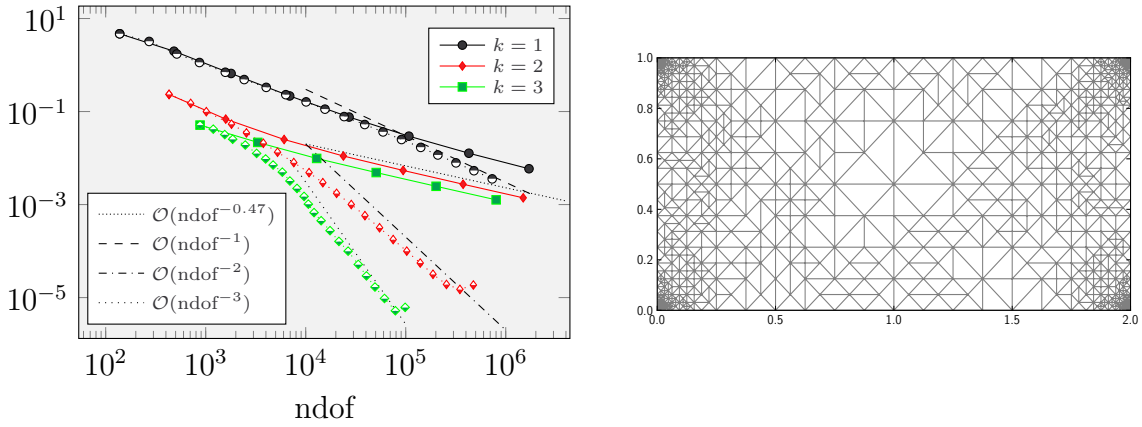


Figure 4.1: Convergence history plot of the relative error $(C^2_{\mathrm{LLB},h} - C^2_{\mathrm{LBB}})/C^2_{\mathrm{LBB}}$ with uniform (solid line) and adaptive (dotted line) mesh refinements and the adaptively refined mesh for $k = 2$ and ndof $= 26996$ in Experiment 2

There exists a function $z^\perp \in Z^\perp \cap H^{1+s}(\Omega;\mathbb{R}^2)\setminus\{0\}$ with $C_{\mathrm{LBB}}\|\nabla z^\perp\|_{L^2(\Omega)} = \|\operatorname{div} z^\perp\|_{L^2(\Omega)}$ and $s = 0.4760291$ [CCDL15, Eq. 3.2]. The (approximated) LBB constant $C^2_{\mathrm{LLB}} = 0.1499719$ [BCDG16, Sec. 5.4.2]. This experiment computes the smallest eigenvalue $\mu_{h,1}$ and the corresponding eigenfunction $\psi_{h,1} \in X_h := S^k_0(\mathcal{T};\mathbb{R}^2) \times (RT_{k-1}(\mathcal{T};\mathbb{R}^{2\times 2}) \cap \Sigma)$, $k = 1,2,3$, with (4.11). The eigenvalue $\mu_{h,1}$ results in the approximation of the LBB constant $C^2_{\mathrm{LBB},h} := 2(1-(1-\mu_{h,1})^2)$. Figure 4.1 displays the convergence history plot of the relative error $(C^2_{\mathrm{LLB},h} - C^2_{\mathrm{LBB}})/C^2_{\mathrm{LBB}}$ for uniformly and adaptively refined meshes $\mathcal{T}$. The adaptive mesh refinement utilizes Algorithm 3 on page 137 with bulk parameter $\Theta = 0.3$ and refinement indicator (4.12) with $\psi_{h,1} = (v_h, \xi_h) \in X_h$. The uniform refinement results in the expected (see Theorem 4.1.2) speed of convergence $(C^2_{\mathrm{LLB},h} - C^2_{\mathrm{LBB}})/C^2_{\mathrm{LBB}} = \mathcal{O}(\text{ndof}^{-0.47})$ with ndof $= \dim X_h$. The adaptive mesh refinement results in strong refinements of the corners (see Figure 4.1) and improves the convergence rate significantly: the experiment with adaptively refined meshes suggests the optimal convergence speed $(C^2_{\mathrm{LLB},h} - C^2_{\mathrm{LBB}})/C^2_{\mathrm{LBB}} = \mathcal{O}(\text{ndof}^{-k})$ for polynomial degrees $k = 1,2,3$. Figure 4.3 indicates that the error indicator $\sum_{T\in\mathcal{T}} \eta^2(T)$ is equivalent to the relative error $(C^2_{\mathrm{LLB},h} - C^2_{\mathrm{LBB}})/C^2_{\mathrm{LBB}}$, that is, the error indicator decreases with the same rate as the error in the eigenvalue approximation. Numerical difficulties cause the failure of the adaptive method for $k = 2$, ndof > 348232 and $k = 3$, ndof > 78794.

Experiment 3 (Essential spectrum). Let $\Omega = (0,1)^2$ be the unit square domain. The essential spectrum of the Cosserat operator equals $[1/2 - 1/\pi, 1/2 + 1/\pi] \cup \{1\}$ [CCDL15, Thm. 3.3] and it is conjectured [CD15, p. 897] that $C^2_{\mathrm{LBB}} = 1/2 - 1/\pi$ is the lower bound of this essential spectrum. This experiment computes the smallest eigenvalues $\mu_{h,1}$ and the corresponding eigenfunction $\psi_{h,1} \in X_h = S^k_0(\mathcal{T};\mathbb{R}^2) \times (RT_{k-1}(\mathcal{T};\mathbb{R}^{2\times 2}) \cap \Sigma)$, $k = 1,2,3$, with (4.11). This computation results in the approximation $C^2_{\mathrm{LBB},h} \searrow C^2_{\mathrm{LBB}}$ from Theorem 4.1.1*(i)*. Figure 4.2 displays the convergence history plot for the relative error $(C^2_{\mathrm{LLB},h} - C^2_{\mathrm{LBB}})/C^2_{\mathrm{LBB}}$. It indicates $(C^2_{\mathrm{LLB},h} - C^2_{\mathrm{LBB}})/C^2_{\mathrm{LBB}} = \mathcal{O}(\text{ndof}^{-0.16})$ for uniformly refined meshes $\mathcal{T}$. The rate is similar to the rate $-1/7$ from [Gal19, Sec. 5.4]. The

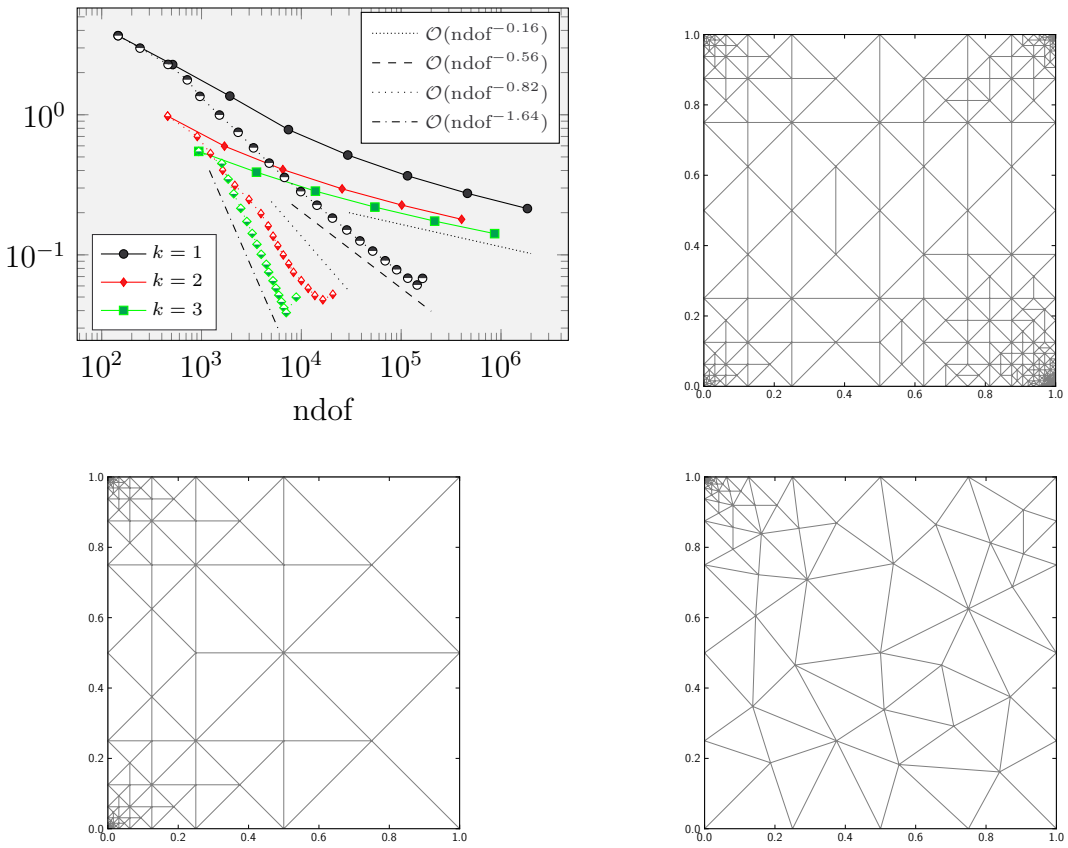


Figure 4.2: Convergence history plot of the relative error $(C^2_{\mathrm{LLB},h} - C^2_{\mathrm{LBB}})/C^2_{\mathrm{LBB}}$ with uniform (solid) and adaptive (dotted) mesh refinements and the adaptively refined meshes for $k = 2$ and ndof $= 13086$ (top right), $k = 3$ and ndof $= 11474$ (bottom left), and $k = 3$ and ndof $= 8199$ (bottom right) with different initial triangulations in Experiment 3

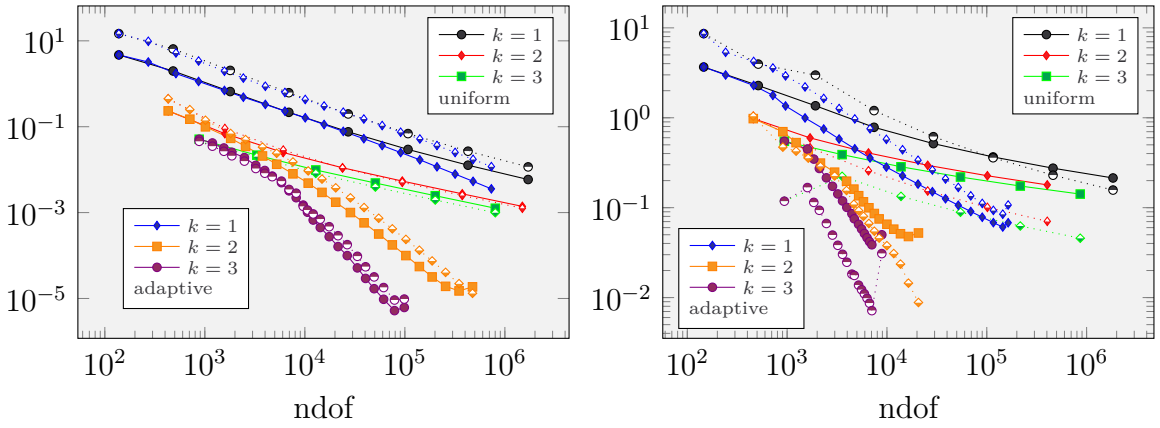


Figure 4.3: Comparison of the relative error $(C^2_{\mathrm{LLB},h} - C^2_{\mathrm{LBB}})/C^2_{\mathrm{LBB}}$ (filled markers) and the error indicator $\eta(\mathcal{T})^2 := 100 \sum_{T\in\mathcal{T}} \eta^2(T)$ (half-filled markers) in Experiment 2 (left) and Experiment 3 (right)

adaptive mesh refinement with Algorithm 3 on page 137 with bulk parameter $\Theta = 0.3$ and refinement indicator (4.12) with $\psi_{h,1} = (v_h, \xi_h)$ improves the rate of convergence. However, the rate is much smaller than in Experiment 2. This indicates that C^2_{LBB} belongs to the essential spectrum. Unlike Experiment 2, the error indicator $\sum_{T\in\mathcal{T}} \eta^2(T)$ converges faster than the error $(C^2_{\mathrm{LLB},h} - C^2_{\mathrm{LBB}})/C^2_{\mathrm{LBB}}$ (see Figure 4.3). Moreover, the adaptively refined meshes depend very much on the initial mesh: Figure 4.2 displays refinements at either one or multiply corners with the adaptive algorithm and different initial triangulations for polynomial degrees $k = 2, 3$. Similar phenomena are well understood for clustered eigenvalues of compact operators [Gal14b, Gal15].

Discussion. The overall conclusion from the numerical benchmarks in this section are in agreement with the theoretical predictions of Theorem 4.1.1–4.1.2. The convergence rates and errors are similar to the results from [Gal19]. The algorithm struggles for moderate numbers of degrees of freedom, that is, numerical difficulties cause large errors in the computation of the smallest eigenvalue $\mu_{h,1}$ with (4.11). The numerical experiments in [Gal19] allow for more degrees of freedom and so result in better approximations of C^2_{LBB}. Either the algorithm from [Gal19] causes less numerical difficulties or the MATLAB implementation in [Gal19] is more robust than the FEniCS implementation of this thesis (see Appendix A.1.2 for a precise description of the implemented routine and the choice of parameters). Thus, it is unclear if the algorithm in [Gal19] performs better. Maybe, a more intense utilization of the Rayleigh quotient structure (4.1) (for example with the application of general Rayleigh quotient iterations [Gel81]) improves the computation of $C^2_{\mathrm{LBB},h}$ and so results in a more robust approximation of C^2_{LBB} than the discrete problem [Gal19, Eq. 8], which solves a mixed eigenvalue problem.

5 Discontinuous Petrov-Galerkin method

The DPG method is a novel approach to approximate the solution to a variational problem with Hilbert spaces X and Y, a functional $F \in Y^*$ in the dual space of Y, and a bilinear from $b : X \times Y \to \mathbb{R}$. More precisely, the method approximates the solution $\mathfrak{u} \in X$ to

$$b(\mathfrak{u}, y) = F(y) \qquad \text{for all } y \in Y. \tag{5.1}$$

The DPG method bases on the idea of optimal test functions by Leszek F. Demkowicz and Jay Gopalakrishnan [DG10, DG11], which reads: Given a discrete space $X_h \subset X$ and an inner product $(\bullet, \bullet)_Y$ in Y, set the trial-to-test operator $T : X \to Y$ with $(Tx, y)_Y = b(x, y)$ for all $(x, y) \in X \times Y$ and seek the solution $\mathfrak{u}_h^i \in X_h$ to

$$b(\mathfrak{u}_h^i, y_h) = F(y_h) \qquad \text{for all } y_h \in TX_h. \tag{5.2}$$

This ansatz ensures that the stability of the continuous problem implies discrete stability in the sense that the inf-sup condition (2.10) implies the discrete inf-sup condition

$$0 < \beta := \inf_{x \in X \setminus \{0\}} \sup_{y \in Y \setminus \{0\}} \frac{b(x, y)}{\|x\|_X \|y\|_Y} \leq \beta_h := \inf_{x_h \in X_h \setminus \{0\}} \sup_{y_h \in TX_h \setminus \{0\}} \frac{b(x_h, y_h)}{\|x_h\|_X \|y_h\|_Y}.$$

This feature is a huge advantage, especially for problems where mixed and Galerkin schemes struggle with discrete stability, for example singularly perturbed problems and problems in fluid dynamics, acoustics, or electrodynamics. On the other hand, the computation of optimal test functions is costly and fails for many problems. The practical DPG method [GQ14] remedies these difficulties. It defines a discrete trial-to-test operator $T_h : X \to Y_h$ with $(T_h x, y_h)_Y = b(x, y_h)$ for all $x \in X$ and y_h in a subspace $Y_h \subset Y$ with dimension $\dim X_h \ll \dim Y_h$ and seeks the solution $\mathfrak{u}_h^p \in X_h$ to

$$b(\mathfrak{u}_h^p, y_h) = F(y_h) \qquad \text{for all } y_h \in T_h X_h. \tag{5.3}$$

Since the computation of the solution $\mathfrak{u}_h^p \in X_h$ to (5.3) is only practicable if the computation of the discrete space $T_h X_h \subset Y_h$ is fast, DPG methods utilize discontinuous test spaces Y. This allows the local (and so highly parallelizable) computation of the discrete trial-to-test operator T_h.

Section 5.1.1 introduces three hypotheses (**H1**)–(**H3**) which imply well-posedness, a priori estimates, and a posteriori error control for DPG methods. The design of variational formulations with broken test spaces Y leads to bilinear forms which decompose into two components. Section 5.1.2 investigates bilinear forms of this form. One of the two components involves traces. Since traces can be very complicated, Section 5.1.3 exemplifies

the design by its application to a problem, where the resulting traces are well understood, namely the Poisson model problem. Section 5.1.4 generalizes the design for a huge class of problems. Thereby, it circumvents the complicated analysis of traces by an alternative definition of trace operators. This alternative definition allows for an analysis with simple functional analytical tools and so leads to well-posed DPG formulations. Section 5.1.5 investigates these DPG formulations and shows a relation of the DPG method and the LSFEM. Section 5.1.6 concludes the analysis of the DPG method with a special case of the abstract framework from Section 5.1.4, namely the ultra-weak DPG method.

Section 5.2 exemplifies the strength of the abstract framework of Section 5.1.4–5.1.6. The first example is the proof of the asymptotic exactness and best approximation result for a primal DPG formulation for the Helmholtz equation in Section 5.2.1. The proof bases on the relation of the DPG method and the LSFEM from Section 5.1.5. Section 5.2.2 shows that knowledge about LSFEMs facilitates the design of DPG methods. More precisely, due to the relation of the DPG method and the LSFEM, knowledge about a locking-free LSFEM for linear elasticity allows to design a locking-free primal DPG method.

5.1 Analysis of DPG

The analysis of this chapter focuses on real-valued Hilbert spaces. However, all results extend to complex-valued Hilbert spaces as well.

5.1.1 Idealized and practical DPG

The functional analytical framework of the DPG method is simple. It bases on three hypotheses (**H1**)–(**H3**) and allows for the proof of well-posedness and quasi-optimality [GQ14]. Moreover, it leads to a built-in a posteriori error control [CDG14]. This section introduces these results briefly. The new contributions of this section are Theorem 5.1.4 and 5.1.9. Theorem 5.1.4 improves the a posteriori error control and Theorem 5.1.9 investigates the error $\|\mathfrak{u}_h^i - \mathfrak{u}_h^p\|_X$ of the solution $\mathfrak{u}_h^i \in X_h$ and $\mathfrak{u}_h^p \in X$ to the idealized (5.2) and practical (5.3) DPG method.

Let X and Y be Hilbert spaces with norms $\|\bullet\|_X$ and $\|\bullet\|_Y$ and discrete subspaces $X_h \subset X$ and $Y_h \subset Y$. The norm in the dual space Y^* of Y reads $\|\bullet\|_{Y^*}$. Let $b : X \times Y \to \mathbb{R}$ be a bilinear form, which satisfies the following three hypotheses.

(**H1**) It holds the uniqueness condition

$$\{x \in X \mid b(x, y) = 0 \text{ for all } y \in Y\} = \{0\}.$$

(**H2**) The bilinear form $b : X \times Y \to \mathbb{R}$ satisfies

$$0 < \beta := \inf_{x \in X \setminus \{0\}} \frac{\|b(x, \bullet)\|_{Y^*}}{\|x\|_X} \leq \|b\| := \sup_{x \in X \setminus \{0\}} \frac{\|b(x, \bullet)\|_{Y^*}}{\|x\|_X} < \infty. \tag{5.4}$$

(**H3**) There exists a linear operator $P : Y \to Y_h$ with, for all $x_h \in X_h$ and $y \in Y$,

$$b(x_h, y - Py) = 0 \qquad \text{and} \qquad \|P\| := \sup_{y \in Y \setminus \{0\}} \|Py\|_Y / \|y\|_Y < \infty. \tag{5.5}$$

The combination of (**H1**)–(**H2**) with the Babuška-Lax-Milgram Theorem (Theorem 2.2.3) proves the existence of unique solutions $\mathfrak{u} \in X$ and $\mathfrak{u}_h^i \in X_h$ to the variational problem (5.1) and the idealized DPG method (5.2). Hypothesis (**H3**) allows for the following result.

Theorem 5.1.1 (A priori analysis)**.** *The hypotheses* (**H1**)–(**H3**) *imply the existence of a unique solution* $\mathfrak{u}_h \in X_h$ *to the practical DPG method (5.3) and*

$$\|\mathfrak{u} - \mathfrak{u}_h\|_X \leq \|P\| \, \|b\| \, \beta^{-1} \min_{x_h \in X_h} \|\mathfrak{u} - x_h\|_X.$$

Proof. This theorem is proven in [GQ14, Thm. 2.1]. □

Remark 5.1.2 (Discrete inf-sup condition)**.** *Suppose* (**H1**)–(**H2**)*, then* (**H3**) *is equivalent to the discrete inf-sup condition [CH16, Lem. 10]*

$$0 < \beta_h := \inf_{x_h \in X_h \setminus \{0\}} \frac{\|b(x_h, \bullet)\|_{Y_h^*}}{\|x_h\|_X}.$$

Thus, the application of Theorem 5.1.1 requires either the design of an operator $P : Y \to Y_h$ *with* (**H3**) *(as for example in [GQ14] and [CDG16]) or the verification of the discrete inf-sup condition (as for example in [CGHW14] and [CH16]).*

Remark 5.1.3 (Instant stability)**.** *The design of linear and bounded operators* $P_T : Y|_T \to Y_h|_T$ *for all* $T \in \mathcal{T}$ *with a local annulation property leads in [GQ14, Lem. 3.1–3.3] and [CDG16, Sec. 5] to an operator* P *with (5.5). This ansatz requires a large discrete space* Y_h *(the polynomial degree of* Y_h *equals the space dimension* d *plus the polynomial degree of* X_h*), but applies to a huge class of problems. Moreover, the constant* $\|P\|$ *is independent of the mesh-size and so implies stability of the DPG method even on coarse meshes, see for example Experiment 4 in Section 5.2.1.*

Besides the a priori estimate in Theorem 5.1.1, (**H1**)–(**H3**) imply the following built-in a posteriori error control.

Theorem 5.1.4 (A posteriori error control)**.** *Under the assumptions of Theorem 5.1.1, the computable residual* $\|b(x_h, \bullet) - F\|_{Y_h^*}$ *satisfies*

$$\beta \|\mathfrak{u} - x_h\|_X \leq \|P\| \, \|b(x_h, \bullet) - F\|_{Y_h^*} + \|F \circ (1 - P)\|_{Y^*} \qquad \text{for all } x_h \in X_h. \tag{5.6}$$

The upper bound is efficient in the sense that $\|b(x_h, \bullet) - F\|_{Y_h^*} \leq \|b\| \, \|\mathfrak{u} - x_h\|_X$ *and* $\|F \circ (1 - P)\|_{Y^*} \leq \|b\| \, \|1 - P\|_{L(Y;Y)} \min_{\xi_h \in X_h} \|\mathfrak{u} - \xi_h\|_X$.

Proof. Given a function $x_h \in X_h$, the inf-sup condition (5.4) and the annulation property (5.5) imply the existence of a function $y \in Y$ with $\|y\|_Y = 1$ and

$$\begin{aligned} \beta \|\mathfrak{u} - x_h\|_X &\leq b(\mathfrak{u} - x_h, y) = F(y - Py) - b(x_h, y - Py) + F(Py) - b(x_h, Py) \\ &\leq \|F \circ (1 - P)\|_{Y^*} + \|P\| \, \|b(x_h, \bullet) - F\|_{Y_h^*}. \end{aligned}$$

The efficiency is proven in [CDG14, Thm. 2.1]. □

Remark 5.1.5 (Error control in [CDG14])**.** *Theorem 5.1.4 improves the estimate in [CDG14, Eq. 2.1], which reads, for all* $x_h \in X_h$*,*

$$\beta \|\mathfrak{u} - x_h\|_X \leq (\|b(x_h, \bullet) - F\|_{Y_h^*}^2 + (\|P\| \, \|b(x_h, \bullet) - F\|_{Y_h^*} + \|F \circ (1 - P)\|_{Y^*})^2)^{1/2}.$$

The following well-known alternative characterizations (see for example [CDW12]) of the solution to the DPG method are important tools in the following proofs. Recall the trial-to-test operator $T_h : X \to Y_h$ with

$$b(x, y_h) = (T_h x, y_h)_Y \qquad \text{for all } x \in X \text{ and } y \in Y. \tag{5.7}$$

Theorem 5.1.6 (DPG as Galerkin FEM)**.** *The function $\mathfrak{u}_h \in X_h$ solves the practical DPG method (5.3) (or the idealized DPG method (5.2) with $Y_h = Y$) if and only if*

$$(T_h \mathfrak{u}_h, T_h x_h)_Y = F(T_h x_h) \qquad \textit{for all } x_h \in X_h. \tag{5.8}$$

Proof. The identity $b(\mathfrak{u}_h, y_h) = (T_h \mathfrak{u}_h, y_h)_Y$ for all $y_h \in Y_h$ yields the equivalence of (5.3) and

$$(T_h \mathfrak{u}_h, T_h x_h)_Y = F(T_h x_h) \qquad \text{for all } x_h \in X_h. \qquad \square$$

Let the inner product $(\bullet, \bullet)_Y$ induce the norm $\|\bullet\|_Y$ in the Hilbert space Y.

Theorem 5.1.7 (DPG as minimal residual method)**.** *The function $\mathfrak{u}_h \in X_h$ solves the practical DPG method (5.3) (or the idealized DPG method (5.2) with $Y_h = Y$) if and only if*

$$\mathfrak{u}_h = \operatorname*{arg\,min}_{x_h \in X_h} \|b(x_h, \bullet) - F\|_{Y_h^*}.$$

Proof. Lemma 3.1.5 and $(T_h \mathfrak{u}, y_h)_Y = F(y_h)$ for all $y_h \in Y_h$ with the trial-to-test operator from (5.7) result in the equivalence of (5.8) and $\mathfrak{u}_h = \arg\min_{x_h \in X_h} \|T_h x_h - T_h \mathfrak{u}\|_Y$. The identity $\|T_h x_h - T_h \mathfrak{u}\|_Y = \|b(x_h, \bullet) - F\|_{Y_h^*}$ for all $x_h \in X_h$ concludes the proof. $\square$

Theorem 5.1.8 (DPG as mixed problem)**.** *The function $\mathfrak{u}_h \in X_h$ solves the practical DPG method (5.3) (or the idealized DPG method (5.2) with $Y_h = Y$) if and only if $(\eta_h, \mathfrak{u}_h) \in Y_h \times X_h$ solves*

$$(\eta_h, y_h)_Y - b(\mathfrak{u}_h, y_h) = -F(y_h) \qquad \textit{for all } y_h \in Y_h, \tag{5.9a}$$

$$b(x_h, \eta_h) = 0 \qquad \textit{for all } x_h \in X_h. \tag{5.9b}$$

Proof. Step 1 ((5.3) $\Longrightarrow$ (5.9)). Given $F \in Y^*$, let $\mathfrak{u}_h \in X_h$ solve (5.3) and define the (unique) Riesz representation $\eta_h \in Y_h$ with (5.9a). Then (5.3) and the definition of the trial-to-test operator $T_h : X \to Y_h$ in (5.7) show

$$b(x_h, \eta_h) = (T_h x_h, \eta_h)_Y = b(\mathfrak{u}_h, T_h x_h) - F(T_h x_h) = 0 \qquad \text{for all } x_h \in X_h.$$

Step 2 ((5.9) $\Longrightarrow$ (5.3)). Recall the trial-to-test operator $T_h : X \to Y_h$ with (5.7) and let $(\eta_h, \mathfrak{u}_h) \in Y_h \times X_h$ solve (5.9). It holds, for all $x_h \in X_h$,

$$b(\mathfrak{u}_h, T_h x_h) = F(T_h x_h) - (\eta_h, T_h x_h)_Y = F(T_h x_h) - b(x_h, \eta_h) = F(T_h x_h). \qquad \square$$

The remainder of this section compares the solution $\mathfrak{u}_h^i = \arg\min_{x_h \in X_h} \|b(x_h, \bullet) - F\|_{Y^*}$ to the idealized (5.2) DPG method and the solution $\mathfrak{u}_h^p = \arg\min_{x_h \in X_h} \|b(x_h, \bullet) - F\|_{Y_h^*}$ to the practical (5.3) DPG method.

Theorem 5.1.9 (Comparison of idealized and practical DPG)**.** *The solutions* $\mathfrak{u}_h^i \in X_h$ *and* $\mathfrak{u}_h^p \in X_h$ *to the idealized (5.2) and the practical (5.3) DPG method satisfy*

$$\beta^2\|\mathfrak{u}_h^i - \mathfrak{u}_h^p\|_X^2 \leq \|b(\mathfrak{u}_h^i - \mathfrak{u}_h^p, \bullet)\|_{Y^*}^2 = \|b(\mathfrak{u}_h^p, \bullet) - F\|_{Y^*}^2 - \|b(\mathfrak{u}_h^i, \bullet) - F\|_{Y^*}^2 \tag{5.10a}$$

$$\leq \|b(\mathfrak{u}_h^p, \bullet) - F\|_{Y^*}^2 - \|b(\mathfrak{u}_h^p, \bullet) - F\|_{Y_h^*}^2. \tag{5.10b}$$

Proof. Define the trial-to-test operator $T : X \to Y$ with $(Tx, y)_Y = b(x, y)$ for all $x \in X$ and $y \in Y$. Theorem 5.1.6 (with $Y_h = Y$) implies the Galerkin orthogonality $(T\mathfrak{u}_h^i - T\mathfrak{u}, T(\mathfrak{u}_h^i - \mathfrak{u}_h^p))_Y = 0$ and so the Pythagorean theorem results in

$$\|T\mathfrak{u}_h^p - T\mathfrak{u}\|_Y^2 = \|T\mathfrak{u}_h^i - T\mathfrak{u}\|_Y^2 + \|T\mathfrak{u}_h^i - T\mathfrak{u}_h^p\|_Y^2. \tag{5.11}$$

The identity $\|Tx\|_X = \|b(x, \bullet)\|_{Y^*}$ for all $x \in X$, (5.4), and (5.11) prove

$$\beta^2\|\mathfrak{u}_h^i - \mathfrak{u}_h^p\|_X^2 \leq \|b(\mathfrak{u}_h^i - \mathfrak{u}_h^p, \bullet)\|_{Y^*}^2 = \|b(\mathfrak{u}_h^p, \bullet) - F\|_{Y^*}^2 - \|b(\mathfrak{u}_h^i, \bullet) - F\|_{Y^*}^2. \tag{5.12}$$

The application of the inequality $\|b(\mathfrak{u}_h^p, \bullet) - F\|_{Y_h^*} \leq \|b(\mathfrak{u}_h^i, \bullet) - F\|_{Y_h^*} \leq \|b(\mathfrak{u}_h^i, \bullet) - F\|_{Y^*}$ to (5.12) concludes the proof. □

Set the Riesz representations $\eta = T(\mathfrak{u}_h^p - \mathfrak{u}) \in Y$ and $\eta_h = T_h(\mathfrak{u}_h^p - \mathfrak{u}) \in Y_h$ with

$$(\eta, y)_Y = b(\mathfrak{u}_h^p, y) - F(y) \qquad \text{for all } y \in Y,$$

$$(\eta_h, y_h)_Y = b(\mathfrak{u}_h^p, y_h) - F(y_h) \qquad \text{for all } y_h \in Y_h.$$

Theorem 5.1.9 shows that the error $\beta^2\|\mathfrak{u}_h^i - \mathfrak{u}_h^p\|_X^2 \leq \|\eta\|_Y^2 - \|\eta_h\|_Y^2 = \|\eta - \eta_h\|_Y^2 = \min_{y_h \in Y_h}\|\eta - y_h\|_Y^2$. Often, the polynomial degree of Y_h is larger than the polynomial degree of X_h. This can lead to an error $\|\mathfrak{u}_h^i - \mathfrak{u}_h^p\|_X$ of higher order (see for example Experiment 1 in Section 5.2.1), that is, there exist constants $C(h) > 0$ with

$$\beta^2\|\mathfrak{u}_h^i - \mathfrak{u}_h^p\|_X^2 \leq \|\eta - \eta_h\|_Y^2 \leq C(h)\|\mathfrak{u} - \mathfrak{u}_h^p\|_X^2 \qquad \text{and} \qquad C(h) < \infty \text{ as } h \to 0. \tag{5.13}$$

5.1.2 Broken variational formulation

Section 5.1.1 introduces some striking advantages of the DPG method like instant stability and built-in error control. However, the practicability relies on the fast computation of the discrete trial-to-test operator $T_h : X \to Y_h$ (in other words, it relies on the fast inversion of the Gram matrix $(G_{jk})_{j,k=1,\dots,N} \in \mathbb{R}^{N\times N}$ with a basis $(y_1, y_2, \dots, y_N)$ of Y_h and $G_{jk} = (y_j, y_k)_Y$ for all $j, k = 1, \dots, N = \dim Y_h$). The key idea of the DPG method is the usage of a broken test space Y, that is a space with discontinuities across the interfaces of a given partition $\mathcal{T}$. This discontinuous space allows for the element-wise (and so highly parallelizable) computation of T_h. Carstensen, Demkowicz, and Gopalakrishnan introduce a general design for variational formulations with discontinuous test space Y in [CDG16]. Their design leads to the following abstract setting. Let $X = V \times \Gamma$ and Y be Hilbert spaces with continuous bilinear forms $b : X \times Y \to \mathbb{R}$, $b_0 : V \times Y \to \mathbb{R}$, and $\langle \bullet, \bullet \rangle_{\partial\mathcal{T}} : \Gamma \times Y \to \mathbb{R}$. The norm in X reads $\|(v, t)\|_X = (\|v\|_V^2 + \|t\|_\Gamma^2)^{1/2}$ for all $(v, t) \in X$ with norms $\|\bullet\|_V$ and $\|\bullet\|_\Gamma$ in the Hilbert spaces V and Γ. Let the bilinear form

$$b(v, t; y) = b_0(v, y) + \langle t, y\rangle_{\partial\mathcal{T}} \qquad \text{for all } (v, t) \in X \text{ and } y \in Y.$$

Moreover, let $Y_0 := \{y_0 \in Y \mid \langle t, y_0\rangle_{\partial\mathcal{T}} = 0 \text{ for all } t \in \Gamma\}$ and assume

$$0 < \beta_0 := \inf_{v\in V\setminus\{0\}} \sup_{y_0\in Y_0\setminus\{0\}} \frac{b_0(v,y_0)}{\|v\|_V\|y_0\|_Y}, \tag{5.14a}$$

$$\|b_0\| := \sup_{v\in V\setminus\{0\}} \sup_{y_0\in Y_0\setminus\{0\}} \frac{b_0(v,y_0)}{\|v\|_V\|y_0\|_Y} < \infty, \tag{5.14b}$$

$$0 < \beta_\Gamma := \inf_{t\in\Gamma\setminus\{0\}} \sup_{y\in Y\setminus\{0\}} \frac{\langle t, y\rangle_{\partial\mathcal{T}}}{\|t\|_\Gamma\|y\|_Y}, \tag{5.14c}$$

$$\{0\} = \{y_0 \in Y_0 \mid b_0(v,y_0) = 0 \text{ for all } v \in V\}. \tag{5.14d}$$

Define the constant

$$\beta_{\text{split}} := \frac{\sqrt{2}\beta_0\beta_\Gamma}{\sqrt{\beta_0^2+\beta_\Gamma^2+\|b_0\|^2+\sqrt{(\beta_0^2+\beta_\Gamma^2+\|b_0\|^2)^2-4\beta_0^2\beta_\Gamma^2}}} \tag{5.15}$$

Theorem 5.1.10 (Splitting lemma)**.** *Suppose (5.14), then* $\{y \in Y \mid b(x,y) = 0$ *for all* $x \in X\} = \{0\}$ *and the constant* β_{split} *from (5.15) satisfies*

$$0 < \beta_{\text{split}} \le \beta := \inf_{x\in X\setminus\{0\}} \sup_{y\in Y\setminus\{0\}} \frac{b(x,y)}{\|x\|_X\|y\|_Y}.$$

Proof. This theorem is a special case of [CDG16, Thm. 3.3] with an improved lower bound from [CP18, Thm. 3.3]. □

Given $F \in Y^*$, the combination of the continuity of the bilinear form b, Theorem 5.1.10, and Theorem 2.2.3 proves the existence of a unique solution $\mathfrak{u} = (u,s) \in X$ to the variational problem $b(u,s;y) = F(y)$ for all $y \in Y$. Theorem 5.1.4 shows that the existence of an operator $P : Y \to Y_h$ with (5.5) and discrete subspaces $Y_h \subset Y$ and $X_h = V_h \times \Gamma_h \subset X$ results for all $x_h = (v_h, t_h) \in X_h$ in the guaranteed upper bound

$$\beta\|(u,s)-(v_h,t_h)\|_X \le \|P\|\,\|b(x_h,\bullet)-F\|_{Y_h^*} + \|F\circ(1-P)\|_{Y^*}. \tag{5.16}$$

The remainder of this section computes improved guaranteed upper bounds for the error in V and in Γ. Let the linear operator $P_0 : Y_0 \to Y_h$ with, for all $x_h \in X_h$ and $y_0 \in Y_0$,

$$b(x_h, y_0 - P_0y_0) = 0 \qquad \text{and} \qquad \|P_0\| := \sup_{y_0\in Y_0\setminus\{0\}} \|P_0(y_0)\|_Y/\|y_0\|_Y < \infty. \tag{5.17}$$

Remark 5.1.11 ($P = P_0$)**.** *Let the operator $P : Y \to Y_h$ satisfy (5.5). Then the operator $P_0 := P|_{Y_0}$ satisfies (5.17). However, there might exists operators $P_0 : Y_0 \to Y_h$ with (5.17) and better properties like a smaller norm $\|P_0\| < \|P|_{Y_0}\|$.*

Theorem 5.1.12 (A posteriori error control)**.** *Assume (5.14). Let $\mathfrak{u} = (u,s) \in X$ solve $b(u,s;y) = F(y)$ for all $y \in Y$ and let $x_h = (v_h, t_h) \in X_h$.*

(i) Suppose there exists an operator $P_0 : Y_0 \to Y_h$ with (5.17), then

$$\beta_0\|u-v_h\|_V \le \|P_0\|\,\|b(x_h,\bullet)-F\|_{Y_h^*} + \|F\circ(1-P_0)\|_{Y_0^*}. \tag{5.18}$$

(ii) Suppose there exists an operator $P : Y \to Y_h$ with (5.5), then

$$\beta_\Gamma(1+\beta_0^{-2}\|b_0\|^2)^{-1/2}\|s-t_h\|_\Gamma \le \|P\|\,\|b(x_h,\bullet)-F\|_{Y_h^*} + \|F\circ(1-P)\|_{Y^*}.$$

Proof of (i). Given $x_h = (v_h, t_h) \in X_h$, the inf-sup condition (5.14a) and the annulation property (5.17) imply the existence of a function $y_0 \in Y_0$ with $\|y_0\|_Y = 1$ and

$$\begin{aligned}\beta_0\|u - v_h\|_V &\le b_0(u - v_h, y_0) = F(y_0 - P_0 y_0) - b(x_h, y_0 - P_0 y_0) + F(P_0 y_0) - b(x_h, P_0 y_0)\\ &\le \|F \circ (1 - P_0)\|_{Y_0^*} + \|P_0\|\,\|b(x_h, \bullet) - F\|_{Y_h^*}.\end{aligned}$$

Proof of (ii). Let $x_h = (v_h, t_h) \in X_h$ and set the bilinear form $b_\rho(x, y) := \rho\, b_0(v, y) + \langle t, y\rangle_{\partial\mathcal{T}}$ for all $x = (v, t) \in X$, $y \in Y$, and weights $\rho > 0$. Define the inf-sup constant

$$\beta(\rho) := \inf_{x \in X\setminus\{0\}} \sup_{y \in Y\setminus\{0\}} \frac{b_\rho(x, y)}{\|x\|_X \|y\|_Y} \qquad \text{for all } \rho > 0.$$

Theorem 5.1.10 shows, for all $\rho > 0$,

$$\beta_{\mathrm{split}}(\rho) := \frac{\sqrt{2}\rho\beta_0\beta_\Gamma}{\sqrt{\rho^2(\beta_0^2 + \|b_0\|^2) + \beta_\Gamma^2 + \sqrt{(\rho^2(\beta_0^2 + \|b_0\|^2) + \beta_\Gamma^2)^2 - 4\rho^2\beta_0^2\beta_\Gamma^2}}} \le \beta(\rho).$$

Since $b_\rho(\rho^{-1}v, t; y) = b(v, t; y)$ for all $(v, t) \in X$, $y \in Y$, and $\rho > 0$, Theorem 5.1.4 implies

$$\begin{aligned}\beta_{\mathrm{split}}(\rho)(\rho^{-2}\|u - v_h\|_V^2 + \|s - t_h\|_\Gamma^2)^{1/2} &\le \|b_\rho(\rho^{-1}(u - v_h), s - t_h; \bullet)\|_{Y^*}\\ &= \|b(x_h, \bullet) - F\|_{Y^*} \le \|P\|\,\|b(x_h, \bullet) - F\|_{Y_h^*} + \|F \circ (1 - P)\|_{Y^*}.\end{aligned} \tag{5.19}$$

A calculation shows $\lim_{\rho\to\infty} \beta_{\mathrm{split}}(\rho) = \beta_0\beta_\Gamma(\beta_0^2 + \|b_0\|^2)^{-1/2}$ and so passing to the limit $\rho \to \infty$ in (5.19) concludes the proof. □

Example 5.1.13 (Improved GUBs for Poisson)**.** *Recall the smallest Dirichlet eigenvalue λ_1 from Theorem 2.2.1 of the Laplace operator. A spectral decomposition shows for the primal DPG formulation of the Poisson model problem in Section 5.1.3 that*

$$\begin{aligned}\beta_0 &:= \inf_{v \in H_0^1(\Omega)\setminus\{0\}} \sup_{w \in H_0^1(\Omega)\setminus\{0\}} \frac{(\nabla v, \nabla w)_{L^2(\Omega)}}{\|\nabla v\|_{L^2(\Omega)}(\|\nabla w\|_{L^2(\Omega)}^2 + \|w\|_{L^2(\Omega)}^2)^{1/2})} = (1 + \lambda_1^{-1})^{-1/2}\\ &\le \|b\| := \sup_{v \in H_0^1(\Omega)\setminus\{0\}} \sup_{w \in H_0^1(\Omega)\setminus\{0\}} \frac{(\nabla v, \nabla w)_{L^2(\Omega)}}{\|\nabla v\|_{L^2(\Omega)}(\|\nabla w\|_{L^2(\Omega)}^2 + \|w\|_{L^2(\Omega)}^2)^{1/2})} = 1.\end{aligned}$$

Moreover, the definition of the norm $\|\bullet\|_\Gamma := \|\bullet\|_{\Gamma_{A^*,1}(\partial\mathcal{T})}$ from Corollary 5.1.32 implies $\beta_\Gamma = 1$. The application of these identities to Theorem 5.1.10 leads to the lower bound

$$\beta_{\mathrm{split}} := \frac{\sqrt{2}}{\sqrt{3 + 2\lambda_1^{-1} + \sqrt{5 + 8\lambda_1^{-1} + 4\lambda_1^{-2}}}} \le \beta.$$

This results on the unit square domain $\Omega = (0, 1)^2$ with Dirichlet eigenvalue $\lambda_1 = 2\pi^2$ in

$$\beta_{\mathrm{split}} = 0.607, \qquad \beta_0 = 0.976, \qquad \text{and} \qquad \beta_\Gamma(1 + \beta_0^{-2}\|b_0\|^2)^{-1/2} = 0.69. \tag{5.20}$$

Remark 5.1.14 (GUBs with highly oscillating right-hand sides)**.** *The importance of guaranteed error control led to the design of a posteriori error estimators with efficiency indices close to one for problems where the oscillation of the right-hand side F is of magnitudes*

smaller than the error (see [CM10, CM11, CM13] for computations of efficiency indices for various estimators). A clever design of the operator P_0 in Theorem 5.1.12(i) with sufficiently large test space Y_h might lead in (5.18) to an efficient GUB for problems with highly oscillating right-hand sides F. More precisely, if there exists a good approximation $t_h \in \Gamma_h$ of the trace $t \in \Gamma$ (computed for example with equilibration [BS08]) and a suitable operator P_0, the DPG framework might allow for (highly parallelizable) computations of efficient GUBs for problems with oscillating data.

5.1.3 Variational formulation for Poisson

The paper [CDG16] introduces a design for variational formulations with broken test spaces. The design requires the introduction of traces. At first glance, these traces can be very complicated. Therefore, this section exemplifies the design by its application to a problem where the resulting traces are the well-understood (see for example [GR86, Chap. 1.2] or [McL00, pp. 100–106]) traces of H^1 and $H(\mathrm{div})$ functions. The problem is the Poisson model problem from Section 2.1, which reads: Given a right-hand side $f \in L^2(\Omega)$ and a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, seek the (weak) solution $u \in H^1_0(\Omega)$ to

$$-\Delta u = f. \tag{5.21}$$

Theorem 5.1.15 (Traces of H^1 and $H(\mathrm{div})$ functions)**.** *Let $\omega \subset \Omega$ be a Lipschitz domain with boundary $\partial\omega$. There exist linear operators*

$$\begin{aligned} &\gamma_0^\omega : H^1(\omega) \to \gamma_0^\omega(H^1(\omega)) =: H^{1/2}(\partial\omega) \subset L^2(\partial\omega) \qquad \text{and} \\ &\gamma_\nu^\omega : H(\mathrm{div},\omega) \to \gamma_\nu^\omega(H(\mathrm{div},\omega)) = H^{1/2}(\partial\omega)^* =: H^{-1/2}(\partial\omega). \end{aligned}$$

Their kernels $\ker\gamma_0^\omega = H^1_0(\omega)$ and $\ker\gamma_\nu^\omega = H_0(\mathrm{div},\omega)$ with $H^1_0(\omega)$ and $H_0(\mathrm{div},\omega)$ from (2.5). For all $v \in H^1(\omega)$ and $q \in H(\mathrm{div},\omega)$ the dual pairing satisfies

$$\langle \gamma_\nu^\omega q, \gamma_0^\omega v\rangle_{\partial\omega} := \langle \gamma_\nu^\omega q, \gamma_0^\omega v\rangle_{H^{-1/2}(\partial\omega),H^{1/2}(\partial\omega)} = (\nabla v, q)_{L^2(\omega)} + (v, \mathrm{div}\, q)_{L^2(\omega)}. \tag{5.22}$$

Proof. Theorem 1.5, Theorem 2.5, Equation (2.17), Corollary 2.8, and Theorem 2.6 from [GR86] prove this theorem. □

Remark 5.1.16 (Dual pairing as integral over the boundary)**.** *The Gauss divergence theorem proves, for sufficiently smooth functions $v \in H^1(\omega)$ and $q \in H(\mathrm{div},\omega)$ with Lipschitz domain $\omega \subset \Omega$ and outer unit normal vector $\nu \in \mathbb{R}^d$, that*

$$\langle \gamma_\nu^\omega q, \gamma_0^\omega v\rangle_{\partial\omega} = (\nabla v, q)_{L^2(\omega)} + (v, \mathrm{div}\, q)_{L^2(\omega)} = \int_{\partial\omega} v\, q\cdot\nu \,\mathrm{d}s.$$

Remark 5.1.17 (Characterization of $H^1_0(\omega)$ and $H_0(\mathrm{div},\omega)$)**.** *Let $\omega \subset \Omega$ be a Lipschitz domain. The definitions in (2.4)–(2.5) show that $v \in H^1(\omega)$ and $q \in H(\mathrm{div},\omega)$ if and only if there exist (unique) functions $\nabla v \in L^2(\omega;\mathbb{R}^d)$ and $\mathrm{div}\, q \in L^2(\omega)$ with*

$$\begin{aligned} (v, \mathrm{div}\, r)_{L^2(\omega)} &= -(\nabla v, r)_{L^2(\omega)} \qquad \text{for all } r \in H_0(\mathrm{div},\omega), \\ (w, \mathrm{div}\, q)_{L^2(\omega)} &= -(\nabla w, q)_{L^2(\omega)} \qquad \text{for all } w \in H^1_0(\omega). \end{aligned} \tag{5.23}$$

The combination with $\ker \gamma_0^\omega = H_0^1(\omega)$ *and* $\ker \gamma_\nu^\omega = H_0(\mathrm{div}, \omega)$ *from Theorem 5.1.15 proves that* $v \in H_0^1(\omega)$ *and* $q \in H_0(\mathrm{div}, \omega)$ *if and only if*

$$\begin{aligned} (v, \mathrm{div}\, r)_{L^2(\omega)} &= -(\nabla v, r)_{L^2(\omega)} \qquad \text{for all } r \in H(\mathrm{div}, \omega), \\ (w, \mathrm{div}\, q)_{L^2(\omega)} &= -(\nabla w, q)_{L^2(\omega)} \qquad \text{for all } w \in H^1(\omega). \end{aligned} \tag{5.24}$$

Let $\mathcal{T}$ be a partition of the domain Ω, that is a decomposition of Ω into a finite number of non-empty and disjoint Lipschitz domains such that the closure

$$\overline{\Omega} = \bigcup_{T \in \mathcal{T}} \overline{T}. \tag{5.25}$$

Remark 5.1.18 (Triangulation $\subset$ Partition)**.** *In many applications, the partition* $\mathcal{T}$ *is a regular triangulation of the domain* Ω *into simplices (more precisely, the set of closed elements* $\{\overline{T} \mid T \in \mathcal{T}\}$ *is a regular triangulation). However, the regularity of the triangulation and the shape of the domains* $T \in \mathcal{T}$ *do not effect the analysis, that is, the DPG method allows for meshes with hanging nodes and curved elements. This feature is very attractive for ultra-weak DPG formulations (see Section 5.1.6), since these methods allow for a simple design of discrete spaces* $X_h \subset X$*.*

The multiplication of (5.21) by a broken test function $w^{\mathrm{pw}} \in Y := H^1(\mathcal{T}) := \{v^{\mathrm{pw}} \in L^2(\Omega) \mid v^{\mathrm{pw}}|_T \in H^1(T) \text{ for all } T \in \mathcal{T}\}$, integration over the domain Ω, and the piecewise integration by parts (5.22) lead to the identity

$$\sum_{T \in \mathcal{T}} \Big((\nabla u, \nabla w^{\mathrm{pw}})_{L^2(T)} - \langle \gamma_\nu^T \nabla u, \gamma_0^T w^{\mathrm{pw}} \rangle_{\partial T} \Big) = (f, w^{\mathrm{pw}})_{L^2(\Omega)}. \tag{5.26}$$

Set the piecewise application of the gradient $\nabla_{NC} : H^1(\mathcal{T}) \to L^2(\Omega; \mathbb{R}^d)$ with

$$(\nabla_{NC} w^{\mathrm{pw}})|_T := \nabla(w^{\mathrm{pw}}|_T) \qquad \text{for all } w^{\mathrm{pw}} \in H^1(\mathcal{T}) \text{ and } T \in \mathcal{T}.$$

Define the trace $\gamma_\nu^{\mathcal{T}} : H(\mathrm{div}, \Omega) \to \prod_{T \in \mathcal{T}} H^{-1/2}(\partial T)$ on the skeleton via $\gamma_\nu^{\mathcal{T}} q := (\gamma_\nu^T q|_T)_{T \in \mathcal{T}}$ for all $q \in H(\mathrm{div}, \Omega)$. Moreover, set for all $v_0 \in H^1(\Omega)$, $w^{\mathrm{pw}} \in H^1(\mathcal{T})$, and $t = (t_T)_{T \in \mathcal{T}} \in H^{-1/2}(\partial \mathcal{T}) := \gamma_\nu^{\mathcal{T}} H(\mathrm{div}, \Omega)$ the bilinear forms

$$\langle t, w^{\mathrm{pw}} \rangle_{\partial \mathcal{T}} := \sum\nolimits_{T \in \mathcal{T}} \langle t_T, \gamma_0^T w^{\mathrm{pw}}|_T \rangle_{\partial T}, \tag{5.27a}$$

$$b(v, t; w^{\mathrm{pw}}) := (\nabla v, \nabla_{NC} w^{\mathrm{pw}})_{L^2(\Omega)} - \langle t, w^{\mathrm{pw}} \rangle_{\partial \mathcal{T}}. \tag{5.27b}$$

Then (5.26) equals $b(u, s; w^{\mathrm{pw}}) = (f, w^{\mathrm{pw}})_{L^2(\Omega)}$.

Theorem 5.1.19 (Primal DPG formulation)**.** *The function* $u \in H_0^1(\Omega)$ *is a (weak) solution to (5.21) if and only if* $(u, s) \in H_0^1(\Omega) \times H^{-1/2}(\partial \mathcal{T})$ *with* $s = \gamma_\nu^{\mathcal{T}} \nabla u$ *satisfies*

$$b(u, s; w^{\mathrm{pw}}) = (f, w^{\mathrm{pw}})_{L^2(\Omega)} \qquad \text{for all } w^{\mathrm{pw}} \in H^1(\mathcal{T}). \tag{5.28}$$

The proof of Theorem 5.1.19 utilizes the following lemma.

Lemma 5.1.20 (Traces of global functions)**.** *The bilinear form* $\langle \bullet, \bullet \rangle_{\partial \mathcal{T}}$ *vanishes for test functions in* $H_0^1(\Omega) \subset H^1(\mathcal{T})$*, that is*

$$\langle t, w \rangle_{\partial \mathcal{T}} = 0 \qquad \text{for all } t \in H^{-1/2}(\partial \mathcal{T}) \text{ and } w \in H_0^1(\Omega).$$

Primal	$\underbrace{-\mathrm{div}}_{A^*}\underbrace{\nabla}_{B} u = f$	Ultra-weak	$\underbrace{\begin{pmatrix} 0 & -\mathrm{div} \\ \nabla & -\mathrm{id} \end{pmatrix}}_{A^*} \underbrace{\mathrm{id}}_{B} \begin{pmatrix} u \\ \sigma \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}$
Mixed 1	$\begin{aligned} \overbrace{\mathrm{id}}^{A_1^*} \overbrace{-\mathrm{div}}^{B_1} \sigma &= f \\ \underbrace{\nabla}_{A_2^*} \underbrace{\mathrm{id}}_{B_2} u \underbrace{-\mathrm{id}}_{C_2} \sigma &= 0 \end{aligned}$	Mixed 2	$\begin{aligned} \overbrace{-\mathrm{div}}^{A_1^*} \overbrace{\mathrm{id}}^{B_1} \sigma &= f \\ \underbrace{\mathrm{id}}_{A_2^*} \underbrace{\nabla}_{B_2} u \underbrace{-\mathrm{id}}_{C_2} \sigma &= 0 \end{aligned}$

Table 5.1: Different abstract operators for the Poisson model problem (with identitiy map id) which lead to primal, mixed, and ultra-weak DPG methods

Proof. Let $t \in H^{-1/2}(\partial\mathcal{T})$, then there exists a function $q \in H(\mathrm{div}, \Omega)$ with $\gamma_\nu^{\mathcal{T}} q = t$. The integration by parts formula (5.22) shows

$$\langle t, w\rangle_{\partial\mathcal{T}} = (\nabla w, q)_{L^2(\Omega)} + (w, \mathrm{div}\, q)_{L^2(\Omega)} = \langle \gamma_\nu^\Omega q, \gamma_0^\Omega w\rangle_{\partial\Omega} = 0 \quad \text{for all } w \in H_0^1(\Omega). \qquad \square$$

Proof of Theorem 5.1.19. Let $u \in H_0^1(\Omega)$ solve (5.21). An integration by parts implies

$$\begin{aligned} (f, w^{\mathrm{pw}})_{L^2(\Omega)} &= -(\Delta u, w^{\mathrm{pw}})_{L^2(\Omega)} = \sum_{T\in\mathcal{T}} \Big((\nabla u, \nabla w^{\mathrm{pw}})_{L^2(T)} - \langle \gamma_\nu^T(\nabla u)|_T, \gamma_0^T w^{\mathrm{pw}}\rangle_{\partial T} \Big) \\ &= b(u, \gamma_\nu^{\mathcal{T}} \nabla u; w^{\mathrm{pw}}) \qquad \text{for all } w^{\mathrm{pw}} \in H^1(\mathcal{T}). \end{aligned}$$

Vice versa, Lemma 5.1.20 shows that $(u, s) \in H_0^1(\Omega) \times H^{-1/2}(\partial\mathcal{T})$ with (5.21) solves

$$b(u, s; w) = (\nabla u, \nabla w)_{L^2(\Omega)} = (f, w)_{L^2(\Omega)} \qquad \text{for all } w \in H_0^1(\Omega) \subset H^1(\mathcal{T}).$$

Thus, $u \in H_0^1(\Omega)$ is a weak solution to (5.21). Let the trace in the skeleton $s = (s_T)_{T\in\mathcal{T}}$. A piecewise integration by parts shows

$$\begin{aligned} 0 &= b(u, s; w^{\mathrm{pw}}) - (f, w^{\mathrm{pw}})_{L^2(\Omega)} = (\nabla u, \nabla_{NC} w^{\mathrm{pw}})_{L^2(\Omega)} - \langle s, w^{\mathrm{pw}}\rangle_{\partial\mathcal{T}} + (\Delta u, w^{\mathrm{pw}})_{L^2(\Omega)} \\ &= \langle \gamma_\nu^{\mathcal{T}} \nabla u - s, w^{\mathrm{pw}}\rangle_{\partial\mathcal{T}} = \sum_{T\in\mathcal{T}} \langle \gamma_\nu^T(\nabla u)|_T - s_T, \gamma_0^T w^{\mathrm{pw}}|_T\rangle_{\partial T} \qquad \text{for all } w^{\mathrm{pw}} \in H^1(\mathcal{T}). \end{aligned}$$

This equality and the surjectivity of $\gamma_0^T : H^1(T) \to H^{1/2}(\partial T)$ for all $T \in \mathcal{T}$ conclude the proof. $\square$

5.1.4 Variational formulation for general problems

The previous section shows that the design of DPG formulations requires traces on the skeleton. The investigation of traces is often very difficult, see for example the analysis of traces of $H(\mathrm{curl})$ functions in [BC01, BCS02]. This complicates the design and the application of DPG methods. The appendix of [DGNS17] circumvents the introduction of complicated traces for Friedrichs systems by the definition of the dual pairing (5.27a) via the action of differential operators, that is, it defines the operator $\gamma_\nu^{\mathcal{T}}$ via

$$\langle \gamma_\nu^{\mathcal{T}} q, w^{\mathrm{pw}}\rangle_{\partial\mathcal{T}} := (\nabla_{NC} w^{\mathrm{pw}}, q)_{L^2(\Omega)} + (w^{\mathrm{pw}}, \mathrm{div}\, q)_{L^2(\Omega)} \quad \text{for all } q \in H(\mathrm{div}, \Omega), w^{\mathrm{pw}} \in H^1(\mathcal{T}).$$

This approach leads to well-posed DPG problems. The following sections utilize this idea to design and analyse the DPG method for an abstract linear problem which compiles primal, mixed, and ultra-weak DPG methods. The abstract problem involves Hilbert spaces $H(A^*,\Omega) \subset L^2(\Omega;\mathbb{R}^{m_A\times n_A})$, $H(B,\Omega)$, and $H(C,\Omega)$, a linear differential operator $A^* : H(A^*,\Omega) \to L^2(\Omega;\mathbb{R}^{m_{A^*}\times n_{A^*}})$ and linear operators $B : H(B,\Omega) \to L^2(\Omega;\mathbb{R}^{m_A\times n_A})$, $C : H(C,\Omega) \to L^2(\Omega;\mathbb{R}^{m_{A^*}\times n_{A^*}})$ with bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, and m_A,n_A, m_{A^*},$n_{A^*} \in \mathbb{N}$. Given a right-hand side $f \in L^2(\Omega;\mathbb{R}^{m_{A^*}\times n_{A^*}})$, the abstract problem seeks $(u,\sigma) \in H(B,\Omega)\times H(C,\Omega)$ with boundary conditions and

$$A^*Bu + C\sigma = f. \tag{5.29}$$

Table 5.1 exemplifies the meanings of the abstract operators for the Poisson model problem.

Assumption 5.1.21 (Differential operators)**.** *For any Lipschitz domain $\omega \subset \Omega$ let the domains $H(A,\omega) \subset L^2(\omega;\mathbb{R}^{m_{A^*}\times n_{A^*}})$ and $H(A^*,\omega) \subset L^2(\omega;\mathbb{R}^{m_A\times n_A})$ be vector spaces and let*

$$A_\omega : H(A,\omega) \to L^2(\omega;\mathbb{R}^{m_A\times n_A}) \quad \text{and} \quad A^*_\omega : H(A^*,\omega) \to L^2(\omega;\mathbb{R}^{m_{A^*}\times n_{A^*}})$$

be linear operators. Moreover, assume for all Lipschitz domains $\omega_1 \subset \omega_2 \subset \Omega$ that

$$(A_{\omega_2}v)|_{\omega_1} = A_{\omega_1}v|_{\omega_1} \qquad \text{for all } v \in H(A,\omega_2), \tag{5.30a}$$
$$(A^*_{\omega_2}\vartheta)|_{\omega_1} = A^*_{\omega_1}\vartheta|_{\omega_1} \qquad \text{for all } \vartheta \in H(A^*,\omega_2). \tag{5.30b}$$

The operators A_ω and A^*_ω depend on the underlying Lipschitz domain $\omega \subset \Omega$. In the remaining sections, the underlying domain is clear from the context or negligible due to the identities in (5.30). This motivates a notation without subscript, that is, $A_\omega =: A$ and $A^*_\omega =: A^*$ for all Lipschitz domains $\omega \subset \Omega$.

Assumption 5.1.22 (Characterization of $H(A,\omega)$ and $H(A^*,\omega)$)**.** *For any Lipschitz domain $\omega \subset \Omega$ let*

$$H_0(A,\omega) \subset H(A,\omega) \qquad \text{and} \qquad H_0(A^*,\omega) \subset H(A^*,\omega)$$

be subspaces. Let $v \in H(A,\omega)$ and $\vartheta \in H(A^*,\omega)$ if and only if there exist unique functions $Av \in L^2(\omega;\mathbb{R}^{m_A\times n_A})$ and $A^*\vartheta \in L^2(\omega;\mathbb{R}^{m_{A^*}\times n_{A^*}})$ with

$$(Av,\xi)_{L^2(\omega)} = (v,A^*\xi)_{L^2(\omega)} \qquad \text{for all } \xi \in H_0(A^*,\omega), \tag{5.31a}$$
$$(Aw,\vartheta)_{L^2(\omega)} = (w,A^*\vartheta)_{L^2(\omega)} \qquad \text{for all } w \in H_0(A,\omega). \tag{5.31b}$$

Throughout this section, text boxes exemplify the abstract definitions and results with the (weak) gradient $A_\omega = \nabla$ and the (weak) divergence $A^*_\omega = -\mathrm{div}$ for all Lipschitz domains $\omega \subset \Omega$. Definition 2.1.1 verifies Assumption 5.1.22 with $H_0(A,\omega) = C^\infty_c(\omega;\mathbb{R})$ and $H_0(A^*,\omega) = C^\infty_c(\omega;\mathbb{R}^d)$. Moreover, the extension of the test functions in $C^\infty_c(\omega;\mathbb{R})$ and $C^\infty_c(\omega;\mathbb{R}^d)$ by zero confirms (5.30).

Lemma 5.1.23 (Hilbert spaces $H(A,\omega)$ and $H(A^*,\omega)$)**.** *For all Lipschitz domains $\omega \subset \Omega$ the spaces $H(A,\omega)$ and $H(A^*,\omega)$ are Hilbert spaces with graph norms*

$$\|\bullet\|_{H(A,\omega)} := (\|\bullet\|^2_{L^2(\omega)} + \|A\bullet\|^2_{L^2(\omega)})^{1/2} \quad \text{and} \quad \|\bullet\|_{H(A^*,\omega)} := (\|\bullet\|^2_{L^2(\omega)} + \|A^*\bullet\|^2_{L^2(\omega)})^{1/2}.$$

Proof. Let $\omega \subset \Omega$ be a Lipschitz domain and let $(v_n)_{n\in\mathbb{N}} \subset H(A,\omega)$ be a Cauchy sequence with respect to the graph norm $\|\bullet\|_{H(A,\omega)}$. Thus, $(v_n)_{n\in\mathbb{N}}$ and $(Av_n)_{n\in\mathbb{N}}$ are Cauchy sequences in the Hilbert spaces $L^2(\omega;\mathbb{R}^{m_{A^*}\times n_{A^*}})$ and $L^2(\omega;\mathbb{R}^{m_A\times n_A})$. Therefore, there exist functions $v \in L^2(\omega;\mathbb{R}^{m_{A^*}\times n_{A^*}})$ and $\chi \in L^2(\omega;\mathbb{R}^{m_A\times n_A})$ with $\|v_n - v\|_{L^2(\omega)} \to 0$ and $\|Av_n - \chi\|_{L^2(\omega)} \to 0$ as $n \to \infty$. The strongly convergent sequences satisfy

$$(v, A^*\xi)_{L^2(\omega)} = \lim_{n\to\infty}(v_n, A^*\xi)_{L^2(\omega)} = \lim_{n\to\infty}(Av_n, \xi)_{L^2(\omega)} = (\chi,\xi)_{L^2(\omega)} \quad \text{for all } \xi \in H_0(A^*,\omega).$$

This verifies (5.31a) and so proves $v \in H(A,\omega)$ with $Av = \chi$ and $\|v - v_n\|_{H(A^*,\omega)} \to 0$ as $n \to \infty$. Similar arguments show that $H(A^*,\omega)$ is a Hilbert space. □

Lemma 5.1.23 is well known for $A = \nabla$ and $A^* = -\mathrm{div}$, see for example [BS02, Thm. 1.3.2] and [Mon03, Thm. 3.22].

Lemma 5.1.24 (Trace operator γ_A^ω)**.** *Let $\omega \subset \Omega$ be a Lipschitz domain. Then Assumption 5.1.21–5.1.22 imply the existence of a bounded linear operator γ_A^ω which maps $H(A,\omega)$ onto a subspace of the dual $H(A^*,\omega)^*$ of $H(A^*,\omega)$. The operator reads*

$$\gamma_A^\omega v := (Av, \bullet)_{L^2(\omega)} - (v, A^*\bullet)_{L^2(\omega)} \qquad \text{for all } v \in H(A,\omega). \tag{5.32}$$

Proof. The definition of γ_A^ω in (5.32) implies the linearity of the operator. The application of the Cauchy-Schwarz inequality to (5.32) proves boundedness. □

Let $\omega \subset \Omega$ be a Lipschitz domain. Smooth functions, for example $v \in C^\infty(\overline{\omega}) := \{w|_\omega \mid w \in C_c^\infty(\mathbb{R};\mathbb{R})\} \subset H^1(\omega)$, have boundary values $v|_{\partial\omega} \in L^2(\partial\omega)$. Textbooks like [GR86, Chap. 1.1] and [BS02, Chap. 1.6] define the map $\gamma_0^\omega : H^1(\omega) \to H^{1/2}(\partial\omega) \subset L^2(\partial\omega)$ as a linear and continuous extension of the map $v \mapsto v|_{\partial\omega} \in L^2(\partial\omega)$ with $v \in C^\infty(\overline{\omega})$. This extension allows for the integration by parts formula (5.22). Lemma 5.1.24 defines the trace γ_A^ω in (5.32) via the action of the differential operators $A = \nabla$ and $A^* = -\mathrm{div}$. In other words, Lemma 5.1.24 utilizes the integration by parts formula (5.22) to define γ_A^ω. The following formula relates the trace operators γ_A^ω and γ_0^ω. Let $\nu \in L^2(\partial\Omega;\mathbb{R}^d)$ denote the outer unit normal vector, let the function $v \in H^1(\omega)$, and let the function q be in the (dense) subspace $C^\infty(\overline{\omega};\mathbb{R}^d) := \{r|_\omega \mid r \in C_c^\infty(\mathbb{R};\mathbb{R}^d)\} \subset H(\mathrm{div},\omega)$. Then

$$\int_{\partial\omega} (\gamma_0^\omega v)\, q\cdot\nu \,\mathrm{d}s = \gamma_A^\omega v(q).$$

The design of γ_0^ω extends to Sobolev spaces $W^{s,p}(\omega) := \{v \in L^p(\omega) \mid \partial^\alpha v \in L^p(\omega)$ for all $|\alpha| \le s\}$ with $s \in \mathbb{N}$ and $1 \le p$, but can be very challenging, see for example [BC01, BCS02] for the analysis of traces of $H(\mathrm{curl},\omega)$ functions. The analysis in this thesis utilizes simple functional analysis for Hilbert space. It applies to a large class of differential operators, but not to Banach spaces. Since the DPG method requires Hilbert spaces, this downside is negligible.

Define the space $\Gamma_A(\partial\omega) := \gamma_A^\omega H(A,\omega)$ with Lipschitz domain $\omega \subset \Omega$ and set

$$\|r\|_{\Gamma_A(\partial\omega)} := \inf\{\|v\|_{H(A,\omega)} \mid v \in H(A,\omega) \text{ with } \gamma_A^\omega v = r\} \quad \text{for all } r \in \Gamma_A(\partial\omega). \tag{5.33}$$

If $\|r\|_{\Gamma_A(\partial\omega)} = 0$, there exists an infimizing sequence $(v_n)_{n\in\mathbb{N}} \subset H(A,\omega)$ with $\gamma_A^\omega v_n = r$ for all $n \in \mathbb{N}$ and $\|v_n\|_{H(A,\omega)} \to 0$ as $n \to \infty$. Thus, the Cauchy-Schwarz inequality proves, for all $\vartheta \in H(A^*,\omega)$,

$$|r(\vartheta)| = |\gamma_A^\omega v_n(\vartheta)| \le \|v_n\|_{H(A,\omega)}\|\vartheta\|_{H(A^*,\omega)} \to 0 \qquad \text{as } n \to \infty.$$

This implies $r(\vartheta) = 0$ for all $\vartheta \in H(A^*,\omega)$ and so $r = 0$. Thus, $\|\bullet\|_{\Gamma_A(\partial\omega)}$ is a norm in $\Gamma_A(\partial\omega)$.

Lemma 5.1.25 (Trace operator $\gamma_{A^*}^\omega$)**.** *Let $\omega \subset \Omega$ be a Lipschitz domain. Assumption 5.1.21–5.1.22 imply the existence of a linear operator $\gamma_{A^*}^\omega : H(A^*,\omega) \to \Gamma_{A^*}(\partial\omega) := \Gamma_A(\partial\omega)^*$ with, for all $v \in H(A,\omega)$ and $\vartheta \in H(A^*,\omega)$,*

$$\gamma_{A^*}^\omega\vartheta(\gamma_A^\omega v) := (Av,\vartheta)_{L^2(\omega)} - (v,A^*\vartheta)_{L^2(\omega)} =: \langle\gamma_{A^*}^\omega\vartheta,\gamma_A^\omega v\rangle_{\partial\omega}. \tag{5.34}$$

Proof. Let $\omega \subset \Omega$ be a Lipschitz domain, $\vartheta \in H(A^*,\omega)$, and $t = \gamma_A^\omega v = \gamma_A^\omega w$ with $v, w \in H(A,\omega)$. The linearity of γ_A^ω and the definitions in (5.32) and (5.34) prove

$$\begin{aligned}\langle\gamma_{A^*}^\omega\vartheta,t\rangle_{\partial\omega} &= (Av,\vartheta)_{L^2(\omega)} - (v,A^*\vartheta)_{L^2(\omega)}\\ &= (A(v-w),\vartheta)_{L^2(\omega)} - (v-w,A^*\vartheta)_{L^2(\omega)} + (Aw,\vartheta)_{L^2(\omega)} - (w,A^*\vartheta)_{L^2(\omega)}\\ &= \gamma_A^\omega(v-w)(\vartheta) + (Aw,\vartheta)_{L^2(\omega)} - (w,A^*\vartheta)_{L^2(\omega)} = (Aw,\vartheta)_{L^2(\omega)} - (w,A^*\vartheta)_{L^2(\omega)}.\end{aligned}$$

This proves that $\gamma_{A^*}^\omega\vartheta : \Gamma_A(\partial\omega) \to \mathbb{R}$ is a well-defined linear functional. The Cauchy-Schwarz inequality implies boundedness and so $\gamma_{A^*}^\omega\vartheta \in \Gamma_A(\partial\omega)^*$. □

Remark 5.1.26 (Surjectivity of γ_A^ω)**.** *Let $\omega \subset \Omega$ be a Lipschitz domain. Since the space $\Gamma_A(\partial\omega) := \gamma_A^\omega H(A,\omega)$ is the image of the operator $\gamma_A^\omega : H(A,\omega) \to \Gamma_A(\partial\omega) \subset H(A^*,\omega)^*$ (and so the operator γ_A^ω is a surjective map onto $\Gamma_A(\partial\omega)$), two functions $t_1, t_2 \in \Gamma_{A^*}(\partial\omega) := \Gamma_A(\partial\omega)^*$ are equal if and only if*

$$\langle t_1 - t_2, \gamma_A^\omega v\rangle_{\partial\omega} = 0 \qquad \text{for all } v \in H(A,\omega).$$

Remark 5.1.27 (Surjectivity of $\gamma_{A^*}^\omega$)**.** *Let $\omega \subset \Omega$ be a Lipschitz domain. Lemma 5.1.25 states that the image $\gamma_{A^*}^\omega H(A^*,\omega) \subset \Gamma_A(\partial\omega)^*$ is a subset. Corollary 5.1.32 proves the equality $\gamma_{A^*}^\omega H(A^*,\omega) = \Gamma_A(\partial\omega)^*$. In other words, the operator $\gamma_{A^*}^\omega$ is a surjection onto the dual space $\Gamma_A(\partial\omega)^*$.*

Let $\omega \subset \Omega$ be a Lipschitz domain and $A = \nabla$, $A^* = -\mathrm{div}$. Textbooks like [GR86, Thm. 2.5] define the trace γ_ν^ω as a linear and continuous extension of the mapping $q \mapsto (q\cdot\nu)|_{\partial\omega}$ with outer unit normal vector $\nu \in \mathbb{R}^d$ and smooth functions $q \in C(\overline{\omega};\mathbb{R}^d) := \{r|_\omega \mid r \in C_c(\mathbb{R};\mathbb{R}^d)\} \subset H(\mathrm{div},\omega)$ from $H(\mathrm{div},\omega)$ onto $H^{-1/2}(\partial\omega) := (\gamma_0^\omega H^1(\omega))^*$. Theorem 5.1.15 verifies, for all $q \in H(\mathrm{div},\omega) = H(A^*,\omega)$ and $H^1(\omega) = H(A,\omega)$,

$$\langle\gamma_\nu^\omega q,\gamma_0^\omega v\rangle_{\partial\omega} = \langle\gamma_{A^*}^\omega q,\gamma_A^\omega v\rangle_{\partial\omega}.$$

Let $\omega \subset \Omega$ be a Lipschitz domain. The characterization in Assumption 5.1.22 shows that functions $v_0 \in H_0(A,\omega)$ and $\vartheta_0 \in H_0(A^*,\omega)$ satisfy

$$\begin{aligned}\gamma_A^\omega v_0(\vartheta) = (Av_0,\vartheta)_{L^2(\omega)} - (v_0,A^*\vartheta)_{L^2(\omega)} = 0 \qquad &\text{for all } \vartheta \in H(A^*,\omega),\\ \langle\gamma_{A^*}^\omega\vartheta_0,\gamma_A^\omega v\rangle_{\partial\omega} = (Av,\vartheta_0)_{L^2(\omega)} - (v,A^*\vartheta_0)_{L^2(\omega)} = 0 \qquad &\text{for all } v \in H(A,\omega).\end{aligned}$$

In other words, $\gamma_A^\omega v_0 = 0$ and $\gamma_{A^*}^\omega \vartheta_0 = 0$ with $\Gamma_A(\partial\omega) := \gamma_A^\omega H(A,\omega)$. Thus, the kernels

$$H_0(A,\omega) \subset \ker \gamma_A^\omega \qquad \text{and} \qquad H_0(A^*,\omega) \subset \ker \gamma_{A^*}^\omega. \tag{5.35}$$

Moreover, the definition of the traces proves that any function $v \in L^2(\omega;\mathbb{R}^{m_{A^*}\times n_{A^*}})$ and $\vartheta \in L^2(\omega;\mathbb{R}^{m_A\times n_A})$ with (5.31) satisfies

$$(Av,\xi)_{L^2(\omega)} = (v,A^*\xi)_{L^2(\omega)} \qquad \text{for all } \xi \in \ker\gamma_{A^*}^\omega, \tag{5.36a}$$
$$(Aw,\vartheta)_{L^2(\omega)} = (w,A^*\vartheta)_{L^2(\omega)} \qquad \text{for all } w \in \ker\gamma_A^\omega. \tag{5.36b}$$

For all Lipschitz domains $\omega \subset \Omega$ the combination of (5.35)–(5.36) allows to assume without loss of generality that the kernels of the trace operators γ_A^ω and $\gamma_{A^*}^\omega$ satisfy

$$H_0(A,\omega) = \ker \gamma_A^\omega \qquad \text{and} \qquad H_0(A^*,\omega) = \ker \gamma_{A^*}^\omega. \tag{5.37}$$

Remark 5.1.28 (Characterization of $H_0(A,\omega)$ and $H_0(A^*,\omega)$)**.** *Assumption 5.1.21–5.1.22 and (5.37) imply that $v \in H_0(A,\omega)$ and $\vartheta \in H_0(A^*,\omega)$ if and only if there exist unique functions $Av \in L^2(\omega;\mathbb{R}^{m_A\times n_A})$ and $A^*\vartheta \in L^2(\omega;\mathbb{R}^{m_{A^*}\times n_{A^*}})$ with*

$$(Av,\xi)_{L^2(\omega)} = (v,A^*\xi)_{L^2(\omega)} \qquad \textit{for all } \xi \in H(A^*,\omega), \tag{5.38a}$$
$$(Aw,\vartheta)_{L^2(\omega)} = (w,A^*\vartheta)_{L^2(\omega)} \qquad \textit{for all } w \in H(A,\omega). \tag{5.38b}$$

Let $\omega \subset \Omega$ be a Lipschitz domain. For $A = \nabla$ and $A^* = -\mathrm{div}$ the identity in (5.37) results in $H_0(A,\omega) = H_0^1(\omega)$ and $H_0(A^*,\omega) = H_0(\mathrm{div},\omega)$.

Let $\mathcal{T}$ be a partition of the domain $\Omega \subset \mathbb{R}^d$ into a finite number of non-empty and disjoint Lipschitz domains with (5.25); see Remark 5.1.18 for a discussion on the relation of $\mathcal{T}$ and (regular) triangulations. Define the space of broken test functions

$$Y := H(A,\mathcal{T}) := \{w^{\mathrm{pw}} \in L^2(\Omega;\mathbb{R}^{m_{A^*}\times n_{A^*}}) \mid w^{\mathrm{pw}}|_T \in H(A,T) \text{ for all } T \in \mathcal{T}\}. \tag{5.39}$$

Let the operator $A_{NC} : H(A,\mathcal{T}) \to L^2(\Omega;\mathbb{R}^{m_A\times n_A})$ with

$$(A_{NC}w^{\mathrm{pw}})|_T := A(w^{\mathrm{pw}}|_T) \qquad \text{for all } w^{\mathrm{pw}} \in H(A,\mathcal{T}) \text{ and } T \in \mathcal{T}. \tag{5.40}$$

For all $\vartheta \in H(A^*,\Omega)$ set the trace on the skeleton

$$\gamma_{A^*}^{\mathcal{T}} : H(A^*,\Omega) \to \prod_{T\in\mathcal{T}} \Gamma_{A^*}(\partial T) \qquad \text{with } \gamma_{A^*}^{\mathcal{T}}\vartheta := (\gamma_{A^*}^T\vartheta|_T)_{T\in\mathcal{T}}. \tag{5.41}$$

For all $w^{\mathrm{pw}} \in H(A,\mathcal{T})$ and $(v,\tau,t) \in \mathcal{X} := H(B,\Omega)\times H(C,\Omega)\times\Gamma_{A^*}(\partial\mathcal{T})$ with $t = (t_T)_{T\in\mathcal{T}} \in \Gamma_{A^*}(\partial\mathcal{T}) := \gamma_{A^*}^{\mathcal{T}} H(A^*,\Omega)$ define the bilinear forms

$$\langle t, w^{\mathrm{pw}}\rangle_{\partial\mathcal{T}} := \textstyle\sum_{T\in\mathcal{T}} \langle t_T, \gamma_A^T w^{\mathrm{pw}}|_T\rangle_{\partial T}, \tag{5.42a}$$
$$b(v,\tau,t;w^{\mathrm{pw}}) := (Bv, A_{NC}w^{\mathrm{pw}})_{L^2(\Omega)} - \langle t, w^{\mathrm{pw}}\rangle_{\partial\mathcal{T}} + (C\tau, w^{\mathrm{pw}})_{L^2(\Omega)}. \tag{5.42b}$$

Theorem 5.1.29 (DPG formulation)**.** *Suppose Assumption 5.1.21–5.1.22. The function $(u,\sigma) \in H(B,\Omega)\times H(C,\Omega)$ solves (5.29) if and only if $(u,\sigma,s) \in H(B,\Omega)\times H(C,\Omega)\times \Gamma_{A^*}(\partial\mathcal{T}) \subset \mathcal{X}$ with $s = \gamma_{A^*}^{\mathcal{T}} Bu$ solves*

$$b(u,\sigma,s;w^{\mathrm{pw}}) = (f,w^{\mathrm{pw}})_{L^2(\Omega)} \qquad \textit{for all } w^{\mathrm{pw}} \in H(A,\mathcal{T}). \tag{5.43}$$

The proof of Theorem 5.1.29 utilizes the following generalization of Lemma 5.1.20.

Lemma 5.1.30 (Traces of global functions)**.** *Suppose Assumption 5.1.21–5.1.22 and (5.37). The bilinear form $\langle \bullet, \bullet \rangle_{\partial\mathcal{T}}$ vanishes for functions in $H_0(A,\Omega) \subset H(A,\mathcal{T}) = Y$, that is*

$$\langle t, w \rangle_{\partial\mathcal{T}} = 0 \qquad \text{for all } t \in \Gamma_{A^*}(\partial\mathcal{T}) \text{ and } w \in H_0(A,\Omega).$$

Proof. Let $t \in \Gamma_{A^*}(\partial\mathcal{T})$ and $w \in H_0(A,\Omega)$. There exists a function $\vartheta \in H(A^*,\Omega)$ with $\gamma^{\mathcal{T}}_{A^*}\vartheta = t$. An integration by parts shows, for all $w \in H_0(A,\Omega) \subset H(A,\mathcal{T})$,

$$\langle t, w \rangle_{\partial\mathcal{T}} = (Aw, \vartheta)_{L^2(\Omega)} - (w, A^*\vartheta)_{L^2(\Omega)} = \langle \gamma^{\Omega}_{A^*}\vartheta, \gamma^{\Omega}_{A} w \rangle_{\partial\Omega} = 0. \qquad \square$$

Proof of Theorem 5.1.29. Let $(u,\sigma) \in H(B,\Omega) \times H(C,\Omega)$ solve (5.29), that is

$$(Bu, Aw)_{L^2(\Omega)} + (C\sigma, w)_{L^2(\Omega)} = (f, w)_{L^2(\Omega)} \qquad \text{for all } w \in H_0(A,\Omega). \tag{5.44}$$

An integration by parts implies

$$\begin{aligned} b(u,\sigma,\gamma^{\mathcal{T}}_{A^*}Bu; w^{\mathrm{pw}}) &= \sum_{T\in\mathcal{T}} \Big((Bu, Aw^{\mathrm{pw}})_{L^2(T)} - \langle \gamma^{T}_{A^*}(Bu)|_T, \gamma^{T}_{A} w^{\mathrm{pw}} \rangle_{\partial T} \Big) + (C\sigma, w^{\mathrm{pw}})_{L^2(\Omega)} \\ &= (A^*Bu, w^{\mathrm{pw}})_{L^2(\Omega)} + (C\sigma, w^{\mathrm{pw}})_{L^2(\Omega)} = (f, w^{\mathrm{pw}})_{L^2(\Omega)} \qquad \text{for all } w^{\mathrm{pw}} \in H(A,\mathcal{T}). \end{aligned}$$

Vice versa, Lemma 5.1.30 shows that $(u,\sigma,s) \in H(B,\Omega) \times H(C,\Omega) \times \Gamma_{A^*}(\partial\mathcal{T})$ with (5.43) satisfies

$$b(u,\sigma,s;w) = (Bu, Aw)_{L^2(\Omega)} + (C\sigma, w)_{L^2(\Omega)} = (f, w)_{L^2(\Omega)} \quad \text{for all } w \in H_0(A,\Omega).$$

Thus, (u,σ) solves (5.44). This and an integration by parts show for $s = (s_T)_{T\in\mathcal{T}}$ that

$$\begin{aligned} 0 &= b(u,\sigma,s;w^{\mathrm{pw}}) - (f, w^{\mathrm{pw}})_{L^2(\Omega)} \\ &= (Bu, A_{NC}w^{\mathrm{pw}})_{L^2(\Omega)} + (C\sigma, w^{\mathrm{pw}})_{L^2(\Omega)} - \langle s, w^{\mathrm{pw}} \rangle_{\partial\mathcal{T}} - (A^*Bu + C\sigma, w^{\mathrm{pw}})_{L^2(\Omega)} \\ &= \langle \gamma^{\mathcal{T}}_{A^*}Bu - s, w^{\mathrm{pw}} \rangle_{\partial\mathcal{T}} = \sum_{T\in\mathcal{T}} \langle \gamma^{T}_{A^*}(Bu)|_T - s_T, \gamma^{T}_{A}(w^{\mathrm{pw}}|_T) \rangle_{\partial T} \qquad \text{for all } w^{\mathrm{pw}} \in H(A,\mathcal{T}). \end{aligned}$$

This equality and the surjectivity of γ^{T}_{A} (see Remark 5.1.26) imply $s = \gamma^{\mathcal{T}}_{A^*}Bu$. $\square$

Let the differential operators $A = B = \nabla$ and $A^* = -\mathrm{div}$. Moreover, let $C : \{0\} \to \{0\}$. Then the abstract variational problem (5.43) equals the primal DPG formulation (5.28) for the Poisson model problem.

5.1.5 Extension of traces

The analysis of the variational problem (5.43) requires a more detailed investigation of the traces from Lemma 5.1.24–5.1.25. This section provides this investigation. Recall the definitions from (5.39)–(5.42) and the space $\mathcal{X} := H(B,\Omega) \times H(C,\Omega) \times \Gamma_{A^*}(\partial\mathcal{T})$ with $\Gamma_{A^*}(\partial\mathcal{T}) := \gamma^{\mathcal{T}}_{A^*}H(A^*,\Omega)$. Moreover, let $\rho > 0$ be a positive weight and define the weighted inner product in the Hilbert space (Lemma 5.1.23) $Y = H(A,\mathcal{T})$

$$(\bullet, \bullet)_{Y_\rho} := (\bullet, \bullet)_{L^2(\Omega)} + \rho\, (A_{NC}\bullet, A_{NC}\bullet)_{L^2(\Omega)}. \tag{5.45}$$

Let $(\bullet,\bullet)_{Y_\rho}$ induce the norm $\|\bullet\|_{Y_\rho}$ and set the dual norm $\|F\|_{Y_\rho^*} := \sup\{F(y) \mid y \in Y \text{ and } \|y\|_{Y_\rho} = 1\}$ for all functionals F in the dual space Y^* of Y. Let the weighted inner product $(\bullet,\bullet)_{H(A^*,\omega,\rho)} := (\bullet,\bullet)_{L^2(\omega)} + \rho\,(A^*\bullet, A^*\bullet)_{L^2(\omega)}$ in $H(A^*,\omega)$ induce the norm $\|\bullet\|_{H(A^*,\omega,\rho)}$ for all Lipschitz domains $\omega \subset \Omega$. The Riesz representation theorem implies the existence of the trial-to-test operator $T_\rho : \mathcal{X} \to Y$ with

$$(T_\rho(v,\tau,t), w^{\mathrm{pw}})_{Y_\rho} = b(v,\tau,t;w^{\mathrm{pw}}) \qquad \text{for all } (v,\tau,t) \in \mathcal{X} \text{ and } w^{\mathrm{pw}} \in H(A,\mathcal{T}). \tag{5.46}$$

Theorem 5.1.31 (Extension operator $\mathcal{E}_\rho$)**.** *Suppose Assumption 5.1.21–5.1.22 and let the weight $\rho > 0$. There exists a linear operator $\mathcal{E}_\rho : \mathcal{X} \to H(A^*,\Omega)$ with*

$$\mathcal{E}_\rho(v,\tau,t) := Bv - \rho A_{NC} T_\rho(v,\tau,t) \qquad \textit{for all } (v,\tau,t) \in \mathcal{X}. \tag{5.47}$$

For all $(v,\tau,t) \in \mathcal{X}$, $T \in \mathcal{T}$, and $\xi \in H_0(A^*,T)$ the operator satisfies

(i) $A^*\mathcal{E}_\rho(v,\tau,t) = T_\rho(v,\tau,t) - C\tau$,
(ii) $\gamma^{\mathcal{T}}_{A^*}\mathcal{E}_\rho(v,\tau,t) = t$,
(iii) $\|\mathcal{E}_\rho(0,0,t)\|_{H(A^*,\Omega,\rho)} = \min\{\|\vartheta\|_{H(A^*,\Omega,\rho)} \mid \vartheta \in H(A^*,\Omega) \text{ with } \gamma^{\mathcal{T}}_{A^*}\vartheta = t\}$,
(iv) $\|\mathcal{E}_\rho(v,\tau,0)\|^2_{H(A^*,T,\rho)} \le \|Bv\|^2_{L^2(T)} + \rho\|C\tau\|^2_{L^2(T)}$,
(v) $\|\mathcal{E}_\rho(v,\tau,t)\|^2_{H(A^*,\Omega,\rho)} = \|\mathcal{E}_\rho(0,0,t)\|^2_{H(A^*,\Omega,\rho)} + \|\mathcal{E}_\rho(v,\tau,0)\|^2_{H(A^*,\Omega,\rho)}$,
(vi) $\|b(v,\tau,t;\bullet)\|^2_{Y_\rho^*} = \rho^{-1}\|Bv - \mathcal{E}_\rho(v,\tau,t)\|^2_{L^2(\Omega)} + \|C\tau + A^*\mathcal{E}_\rho(v,\tau,t)\|^2_{L^2(\Omega)}$,
(vii) $\rho\,(A^*\mathcal{E}_\rho(v,\tau,t), A^*\xi)_{L^2(T)} + (\mathcal{E}_\rho(v,\tau,t),\xi)_{L^2(T)} = (Bv,\xi)_{L^2(T)} - \rho\,(C\tau, A^*\xi)_{L^2(T)}$.

Proof of (i). For all $(v,\tau,t) \in \mathcal{X}$ and $\xi \in H_0(A,\Omega) \subset H(A,\mathcal{T}) = Y$ the definition of $\mathcal{E}_\rho$, the identity in (5.46), the identity $A\xi = A_{NC}\xi$ from (5.30a), and Lemma 5.1.30 imply

$$\begin{aligned}(\mathcal{E}_\rho(v,\tau,t), A\xi)_{L^2(\Omega)} &= (Bv, A\xi)_{L^2(\Omega)} - (T_\rho(v,\tau,t),\xi)_{Y_\rho} + (T_\rho(v,\tau,t),\xi)_{L^2(\Omega)}\\ &= (T_\rho(v,\tau,t),\xi)_{L^2(\Omega)} - (C\tau,\xi)_{L^2(\Omega)} + \langle t,\xi\rangle_{\partial\mathcal{T}} = (T_\rho(v,\tau,t) - C\tau,\xi)_{L^2(\Omega)}.\end{aligned}$$

Thus, (5.31b) results in *(i)*.

Proof of (ii). For all $(v,\tau,t) \in \mathcal{X}$ and $w^{\mathrm{pw}} \in H(A,\mathcal{T})$ a piecewise integration by parts (5.34), the definition of $\mathcal{E}_\rho$, *(i)*, and (5.46) show

$$\begin{aligned}\langle \gamma^{\mathcal{T}}_{A^*}\mathcal{E}_\rho(v,\tau,t), w^{\mathrm{pw}}\rangle_{\partial\mathcal{T}} &= (\mathcal{E}_\rho(v,\tau,t), A_{NC}w^{\mathrm{pw}})_{L^2(\Omega)} - (A^*\mathcal{E}_\rho(v,\tau,t), w^{\mathrm{pw}})_{L^2(\Omega)}\\ &= (Bv, A_{NC}w^{\mathrm{pw}})_{L^2(\Omega)} + (C\tau, w^{\mathrm{pw}})_{L^2(\Omega)} - (T_\rho(v,\tau,t), w^{\mathrm{pw}})_{Y_\rho} = \langle t, w^{\mathrm{pw}}\rangle_{\partial\mathcal{T}}.\end{aligned}$$

This and the surjectivity of the trace operator γ^T_A for all $T \in \mathcal{T}$ (see Remark 5.1.26) lead to $\gamma^{\mathcal{T}}_{A^*}\mathcal{E}_\rho(v,\tau,t) = t$.

Proof of (iii). Let $\vartheta \in H(A^*,\Omega)$ with $\gamma^{\mathcal{T}}_{A^*}\vartheta = t \in \Gamma_{A^*}(\partial\mathcal{T})$. Since *(ii)* implies $\gamma^{\mathcal{T}}_{A^*}(\vartheta - \mathcal{E}_\rho(0,0,t)) = 0$, the identity from *(i)*, the definition of $\mathcal{E}_\rho$, and an integration by parts yield

$$\begin{aligned}&(\mathcal{E}_\rho(0,0,t), \vartheta - \mathcal{E}_\rho(0,0,t))_{H(A^*,\Omega,\rho)}\\ &= \rho\,(A^*\mathcal{E}_\rho(0,0,t), A^*(\vartheta - \mathcal{E}_\rho(0,0,t)))_{L^2(\Omega)} + (\mathcal{E}_\rho(0,0,t), \vartheta - \mathcal{E}_\rho(0,0,t))_{L^2(\Omega)}\\ &= \rho\,(T_\rho(0,0,t), A^*(\vartheta - \mathcal{E}_\rho(0,0,t)))_{L^2(\Omega)} - \rho\,(A_{NC}T_\rho(0,0,t), \vartheta - \mathcal{E}_\rho(0,0,t))_{L^2(\Omega)} = 0.\end{aligned} \tag{5.48}$$

Thus, the Pythagorean theorem results in $0 \le \|\mathcal{E}_\rho(0,0,t) - \vartheta\|^2_{H(A^*,\Omega,\rho)} = \|\vartheta\|^2_{H(A^*,\Omega,\rho)} - \|\mathcal{E}_\rho(0,0,t)\|^2_{H(A^*,\Omega,\rho)}$ and concludes the proof of *(iii)*.

Proof of (iv). Property *(ii)* shows $\gamma_{A^*}^{\mathcal{T}}\mathcal{E}_\rho(v,\tau,0)=0$ for all $v\in H(B,\Omega)$ and $\tau\in H(C,\Omega)$. Thus, a piecewise integration by parts, the definition of $\mathcal{E}_\rho$, the identity from *(i)*, and the Cauchy-Schwarz inequality prove, for all $T\in\mathcal{T}$,

$$\begin{aligned}\|\mathcal{E}_\rho(v,\tau,0)\|^2_{H(A^*,T,\rho)} &= \rho\,(T_\rho(v,\tau,0)-C\tau, A^*\mathcal{E}_\rho(v,\tau,0))_{L^2(T)}\\ &\quad+(Bv-\rho A_{NC}T_\rho(v,\tau,0),\mathcal{E}_\rho(v,\tau,0))_{L^2(T)}\\ &= -\rho\,(C\tau,A^*\mathcal{E}_\rho(v,\tau,0))_{L^2(T)}+(Bv,\mathcal{E}_\rho(v,\tau,0))_{L^2(T)}\\ &\le(\rho\|C\tau\|^2_{L^2(T)}+\|Bv\|^2_{L^2(T)})^{1/2}\|\mathcal{E}_\rho(v,\tau,0)\|_{H(A^*,T,\rho)}.\end{aligned}\tag{5.49}$$

Proof of (v). The linearity of $\mathcal{E}_\rho$, Theorem 5.1.31*(ii)*, and (5.48) with $\vartheta:=\mathcal{E}_\rho(v,\tau,t)$ prove $(\mathcal{E}_\rho(0,0,t),\mathcal{E}_\rho(v,\tau,0))_{H(A^*,\Omega,\rho)}=0$ for all $(v,\tau,t)\in\mathcal{X}$. The Pythagorean theorem concludes the proof of *(v)*.

Proof of (vi). Let $(v,\tau,t)\in\mathcal{X}$. The Riesz representation theorem implies

$$\|b(v,\tau,t;\bullet)\|^2_{Y_\rho^*}=\|T_\rho(v,\tau,t)\|^2_{Y_\rho}=\rho\|A^*_{NC}T_\rho(v,\tau,t)\|^2_{L^2(\Omega)}+\|T_\rho(v,\tau,t)\|^2_{L^2(\Omega)}.$$

This, the identity from *(i)*, and the definition of $\mathcal{E}_\rho$ result in *(vi)*.

Proof of (vii). Let $T\in\mathcal{T}$, $\xi\in H_0(A^*,T)$, and $(v,\tau,t)\in\mathcal{X}$. The definition of $\mathcal{E}_\rho$, the identity in *(i)*, and an integration by parts imply

$$\begin{aligned}&\rho\,(A^*\mathcal{E}_\rho(v,\tau,t),A^*\xi)_{L^2(T)}+(\mathcal{E}_\rho(v,\tau,t),\xi)_{L^2(T)}\\ &=\rho\,(T_\rho(v,\tau,t)-C\tau,A^*\xi)_{L^2(T)}+(Bv-\rho A_{NC}T_\rho(v,\tau,t),\xi)_{L^2(T)}\\ &=-\rho\,(C\tau,A^*\xi)_{L^2(T)}+(Bv,\xi)_{L^2(T)}.\end{aligned}$$

□

Suppose the setting of Section 5.1.3, that is, $A=B=\nabla$, $A^*=-\mathrm{div}$, and $C:\{0\}\to\{0\}$. Then the bilinear form $b(v,t;w^{\mathrm{pw}})$ from (5.27b) equals the bilinear form $b(v,0,t;w^{\mathrm{pw}})$ from (5.42b) for all $(v,t)\in X=H^1_0(\Omega)\times\Gamma^*_A(\partial\mathcal{T})$ and $w^{\mathrm{pw}}\in H^1(\mathcal{T})=H(A,\mathcal{T})$. Moreover, Theorem 5.1.31*(vi)* reads (with $\rho=1$)

$$\|b(v,t;\bullet)\|^2_{Y_1^*}=\|\nabla v-\mathcal{E}_1(v,0,t)\|^2_{L^2(\Omega)}+\|\mathrm{div}\,\mathcal{E}_1(v,0,t)\|^2_{L^2(\Omega)}\qquad\text{for all }(v,t)\in X.$$

The right-hand side equals the squared norm $\|(v,\mathcal{E}_1(v,0,t))\|^2_a$ from (3.35).

Theorem 5.1.31 allows for the following generalization of the duality lemma [CDG16, Thm. 2.3]. Set for all $r=(r_T)_{T\in\mathcal{T}}\in\prod_{T\in\mathcal{T}}\Gamma_A(\partial T)$ and $t=(t_T)_{T\in\mathcal{T}}\in\Gamma_{A^*}(\partial\mathcal{T})$ the bilinear form $\langle t,r\rangle_{\partial\mathcal{T}}:=\sum_{T\in\mathcal{T}}\langle t_T,r_T\rangle_{\partial T}$ and the weighted minimal extension norm

$$\|r\|_{\Gamma_{A,\rho}(\partial\mathcal{T})}=\inf\{\|w^{\mathrm{pw}}\|_{Y_\rho}\mid w^{\mathrm{pw}}\in H(A,\mathcal{T})\text{ with }\gamma_A^T w^{\mathrm{pw}}|_T=r_T\text{ for all }T\in\mathcal{T}\}.\tag{5.50}$$

Corollary 5.1.32 (Duality lemma)**.** *For all weights $\rho>0$ and traces $t\in\Gamma_{A^*}(\partial\mathcal{T})$ the minimal extension norm*

$$\begin{aligned}\|t\|_{\Gamma_{A^*,\rho}(\partial\mathcal{T})}&:=\rho^{-1/2}\min\{\|\vartheta\|_{H(A^*,\Omega,\rho)}\mid\vartheta\in H(A^*,\Omega)\text{ and }\gamma_{A^*}^{\mathcal{T}}\vartheta=t\}\\ &=\sup_{w^{\mathrm{pw}}\in Y\setminus\{0\}}\frac{\langle t,w^{\mathrm{pw}}\rangle_{\partial\mathcal{T}}}{\|w^{\mathrm{pw}}\|_{Y_\rho}}=\sup_{r\in(\prod_{T\in\mathcal{T}}\Gamma_A(\partial T))\setminus\{0\}}\frac{\langle t,r\rangle_{\partial\mathcal{T}}}{\|r\|_{\Gamma_{A,\rho}(\partial\mathcal{T})}}.\end{aligned}$$

Proof. The combination of *(ii)*, *(iii)*, and *(vi)* from Theorem 5.1.31, the surjectivity of the operator γ_A^T for all $T \in \mathcal{T}$ (see Remark 5.1.26), and the definition of the minimal extension norm $\|\bullet\|_{\Gamma_{A,\rho}(\partial\mathcal{T})}$ in (5.50) proves, for all weights $\rho > 0$ and traces $t \in \Gamma_{A^*}(\partial\mathcal{T})$,

$$\begin{aligned}\rho^{-1/2}\min\{\|\vartheta\|_{H(A^*,\Omega,\rho)} \mid \vartheta \in H(A^*,\Omega) \text{ and } \gamma_{A^*}^{\mathcal{T}}\vartheta = t\} &= \rho^{-1/2}\|\mathcal{E}_\rho(0,0,t)\|_{H(A^*,\Omega,\rho)}\\ = \|b(0,0,t;\bullet)\|_{Y_\rho^*} = \sup_{w^{\mathrm{pw}}\in Y\setminus\{0\}} \frac{\langle t, w^{\mathrm{pw}}\rangle_{\partial\mathcal{T}}}{\|w^{\mathrm{pw}}\|_{Y_\rho}} &= \sup_{r\in(\prod_{T\in\mathcal{T}}\Gamma_A(\partial T))\setminus\{0\}} \frac{\langle t, r\rangle_{\partial\mathcal{T}}}{\|r\|_{\Gamma_{A,\rho}(\partial\mathcal{T})}}.\end{aligned}$$ □

Remark 5.1.33 (Minimal extension of traces in [CDG16])**.** *Carstensen, Demkowicz, and Gopalakrishnan utilize the variational equation from Theorem 5.1.31(vii) with right-hand side zero to define the minimal extension of the traces for their proof of Corollary 5.1.32.*

Corollary 5.1.34 ($\Gamma_{A^*}(\partial\mathcal{T})$ is a Hilbert space)**.** *The space $\Gamma_{A^*}(\partial\mathcal{T})$ is a Hilbert space with the minimal extension norm $\|\bullet\|_{\Gamma_{A^*,\rho}(\partial\mathcal{T})}$ from Corollary 5.1.32 and weight $\rho > 0$.*

Proof. Corollary 5.1.32 proves that $(\mathcal{E}_\rho(0,0,\bullet), \mathcal{E}_\rho(0,0,\bullet))_{H(A^*,\Omega,\rho)}$ is an inner product in the space $\Gamma_{A^*}(\partial\mathcal{T})$. Let $(t_n)_{n\in\mathbb{N}} \subset \Gamma_{A^*}(\partial\mathcal{T})$ be a Cauchy sequence with respect to the induced norm $\|\bullet\|_{\Gamma_{A^*,\rho}(\partial\mathcal{T})} = \|\mathcal{E}_\rho(0,0,\bullet)\|_{H(A^*,\Omega,\rho)}$. Since $H(A^*,\Omega,\rho)$ is a Hilbert space, there exists a function $\vartheta \in H(A^*,\Omega)$ with $\|\mathcal{E}_\rho(0,0,t_n) - \vartheta\|_{H(A^*,\Omega,\rho)} \to 0$ as $n \to \infty$. Set $t_\infty := \gamma_{A^*}^{\mathcal{T}}\vartheta \in \Gamma_{A^*}(\partial\mathcal{T})$. The definition of $\|\bullet\|_{\Gamma_{A^*,\rho}(\partial\mathcal{T})}$ and Theorem 5.1.31*(ii)* imply

$$\begin{aligned}\|t_n - t_\infty\|_{\Gamma_{A^*,\rho}(\partial\mathcal{T})} &= \|\gamma_{A^*}^{\mathcal{T}}(\mathcal{E}_\rho(0,0,t_n) - \vartheta)\|_{\Gamma_{A^*,\rho}(\partial\mathcal{T})}\\ &\le \|\mathcal{E}_\rho(0,0,t_n) - \vartheta\|_{H(A^*,\Omega,\rho)} \to 0 \qquad \text{as } n\to\infty.\end{aligned}$$ □

Define the space $Z := V \times W$ with closed subspaces $V \subset H(B,\Omega) \times H(C,\Omega)$ and $H_0(A^*,\Omega) \subset W \subset H(A^*,\Omega)$ and set the squared (semi-) norm

$$\|(v,\tau,\vartheta)\|_{a_\rho}^2 := \rho^{-1}\|Bv - \vartheta\|_{L^2(\Omega)}^2 + \|C\tau + A^*\vartheta\|_{L^2(\Omega)}^2 \qquad \text{for all } (v,\tau,\vartheta) \in Z.$$

Given a discrete subspace $X_h \subset X := V \times \gamma_{A^*}^{\mathcal{T}} W$, Theorem 5.1.31*(ii)* proves that

$$Z_h(\rho) := \{(v_h, \tau_h, \mathcal{E}_\rho(v_h,\tau_h,t_h)) \mid (v_h,\tau_h,t_h) \in X_h\} \subset Z$$

is a subspace. Given $f \in L^2(\Omega;\mathbb{R}^{m_{A^*}\times n_{A^*}})$, let $(u,\sigma,s) \in X$ solve the variational problem $b(u,\sigma,s;w^{\mathrm{pw}}) = (f, w^{\mathrm{pw}})_{L^2(\Omega)}$ for all $w^{\mathrm{pw}} \in Y = H(A,\mathcal{T})$.

Theorem 5.1.35 (Idealized DPG as LSFEM)**.** *Suppose Assumption 5.1.21–5.1.22 and let $\rho > 0$. The function $(u_h,\sigma_h,s_h) \in X_h$ solves the idealized DPG method (5.2) if and only if*

$$(u_h,\sigma_h,\mathcal{E}_\rho(u_h,\sigma_h,s_h)) = \operatorname*{arg\,min}_{(v_h,\tau_h,\vartheta_h)\in Z_h(\rho)} \|(u,\sigma,\mathcal{E}_\rho(u,\sigma,s)) - (v_h,\tau_h,\vartheta_h)\|_{a_\rho}.$$

Proof. Theorem 5.1.31*(vi)* and Theorem 5.1.7 prove this theorem. □

Remark 5.1.36 (Practical DPG as LSFEM)**.** *An equivalence of lowest-order practical DPG and LSFEMs is pointed out in [CBHW18, Thm. 3.11] and [Hel18, Chap. 7].*

Remark 5.1.37 (Subspaces V and W)**.** *The definition of subspaces $V \subset H(B,\Omega) \times H(C,\Omega)$ and $H_0(A^*,\Omega) \subset W \subset H(A^*,\Omega)$ allows to include homogeneous boundary conditions and a relation of u and σ.*

The remainder of this section investigates the asymptotic behaviour of the operator $\mathcal{E}_\rho$. The investigation is motivated by two considerations. First, the asymptotic behaviour of $\mathcal{E}_\rho$ enables the possibility to apply the asymptotic exactness results from Section 3.1.1 to DPG methods (see Section 5.2.1). Second, it proves that the following bounds, which result from Theorem 5.1.31*(vi)*, are sharp. Let $\|\bullet\|_V$ be a norm in V and set $\|(v,\tau,t)\|_{X_\rho}^2 := \|(v,\tau)\|_V^2 + \|\mathcal{E}_\rho(v,\tau,t)\|_{H(A^*,\Omega,\rho)}^2$ for all $(v,\tau,t) \in X := V \times \gamma_{A^*}^{\mathcal{T}} W$. Then

$$\beta_\rho^2 := \inf_{(v,\tau,\vartheta)\in Z\setminus\{0\}} \frac{\|(v,\tau,\vartheta)\|_{a_\rho}^2}{\|(v,\tau)\|_V^2 + \rho^{-1}\|\vartheta\|_{H(A^*,\Omega,\rho)}^2} \leq \inf_{(v,\tau,t)\in X\setminus\{0\}} \frac{\|b(v,\tau,t;\bullet)\|_{Y_\rho^*}^2}{\|(v,\tau,t)\|_{X_\rho}^2}, \tag{5.51a}$$

$$\sup_{(v,\tau,t)\in X\setminus\{0\}} \frac{\|b(v,\tau,t;\bullet)\|_{Y_\rho^*}^2}{\|(v,\tau,t)\|_{X_\rho}^2} \leq \sup_{(v,\tau,\vartheta)\in Z\setminus\{0\}} \frac{\|(v,\tau,\vartheta)\|_{a_\rho}^2}{\|(v,\tau)\|_V^2 + \rho^{-1}\|\vartheta\|_{H(A^*,\Omega,\rho)}^2} =: \|b\|_\rho^2. \tag{5.51b}$$

Remark 5.1.38 (Computation of β_ρ and $\|b\|_\rho$)**.** *The techniques from Section 3.2.2–3.2.3 often allow the computation of β_ρ and $\|b\|_\rho$.*

Let $(\mathcal{T}_\ell)_{\ell\in\mathbb{N}}$ be a sequence of partitions of the domain Ω into a finite number of non-empty and disjoint Lipschitz domains with (5.25), that is,

$$\overline{\Omega} = \bigcup_{T\in\mathcal{T}_\ell} \overline{T} \qquad \text{for all } \ell \in \mathbb{N}. \tag{5.52}$$

Let the sequence be nested in the sense that

$$\text{for all } T \in \mathcal{T}_{\ell+1} \text{ and } \ell \in \mathbb{N} \text{ exists a domain } K \in \mathcal{T}_\ell \text{ with } T \subset K. \tag{5.53}$$

Define the ansatz space $X(\mathcal{T}_\ell) := V \times \gamma_{A^*}^{\mathcal{T}_\ell} W$ and the broken test space $Y(\mathcal{T}_\ell) := H(A,\mathcal{T}_\ell)$ with inner product $(\bullet,\bullet)_{Y_\rho(\mathcal{T}_\ell)} := (\bullet,\bullet)_{L^2(\Omega)} + \rho\,(A_{NC}\bullet, A_{NC}\bullet)_{L^2(\Omega)}$ (the notation hides that the operator A_{NC} depends on the triangulation $\mathcal{T}_\ell$ as well), induced norm $\|\bullet\|_{Y_\rho(\mathcal{T}_\ell)}$, and dual space $Y_\rho(\mathcal{T}_\ell)^*$ for all $\ell \in \mathbb{N}$ and $\rho > 0$. Define the bilinear form $b_\ell : X(\mathcal{T}_\ell) \times Y(\mathcal{T}_\ell) \to \mathbb{R}$ as in (5.42b), the trial-to-test operator $T_{\rho,\ell} : X(\mathcal{T}_\ell) \to Y(\mathcal{T}_\ell)$ as in (5.46), and the extension operator $\mathcal{E}_{\rho,\ell} : X(\mathcal{T}_\ell) \to H(A^*,\Omega)$ as in (5.47) for all $\ell \in \mathbb{N}$ and $\rho > 0$. Suppose that

$$\vartheta \in W \text{ with } \gamma_{A^*}^{\mathcal{T}_\ell}\vartheta = 0 \text{ for all } \ell \in \mathbb{N} \text{ implies } \vartheta = 0. \tag{5.54}$$

Lemma 5.1.39 (Operators with (5.54))**.** *Suppose Assumption 5.1.21–5.1.22 and let the space of constant functions be in the domains and kernels of the operators $A : H(A,T) \to L^2(T;\mathbb{R}^{m_A\times n_A})$ for all $T \in \mathcal{T}_\ell$ and $\ell \in \mathbb{N}$, that is*

$$\mathbb{P}_0(T;\mathbb{R}^{m_A\times n_A}) \subset A(H(A,T)) \qquad \text{and} \qquad \mathbb{P}_0(T;\mathbb{R}^{m_{A^*}\times n_{A^*}}) \subset \ker A. \tag{5.55}$$

Moreover, let the ratio of the domains do not degenerate in the sense that there exists a constant $c > 0$ such that for all $T \in \mathcal{T}_\ell$ and $\ell \in \mathbb{N}$ the Lebesgue measure $|B(T)|$ of the smallest ball $B(T) \supset T$ and the Lebesgue measure $|T|$ of T satisfy

$$c|B(T)| \leq |T|. \tag{5.56}$$

Finally, assume that the maximal mesh-size tends to zero, that is,

$$\max\{\operatorname{diam}(T) \mid T \in \mathcal{T}_\ell\} \to 0 \text{ as } \ell \to \infty. \tag{5.57}$$

Then (5.54) holds.

Remark 5.1.40 (Examples for (5.55))**.** *The operators* $A = \nabla$, $A^* = -\mathrm{div}$ *and* $A = \mathrm{curl}$, $A^* = \mathrm{curl}$ *satisfy (5.55) and so (5.54).*

Proof of Lemma 5.1.39. Let $\vartheta \in H(A^*, \Omega)$ with $\gamma_{A^*}^{\mathcal{T}_\ell}\vartheta = 0$ for all $\ell \in \mathbb{N}$. Then the trace $\gamma_{A^*}^{T}\vartheta|_T = 0$ for all $T \in \mathcal{T}_\ell$, $\ell \in \mathbb{N}$. The integration by parts formula (5.34) and $\mathbb{P}_0(T; \mathbb{R}^{m_{A^*} \times n_{A^*}}) \subset \ker A$ result in

$$\int_T A^*\vartheta \,\mathrm{d}x = 0 \qquad \text{for all } \ell \in \mathbb{N} \text{ and } T \in \mathcal{T}_\ell.$$

Since the Lebesgue measure of $|T|$ with $T \in \mathcal{T}_\ell$ vanishes as $\ell \to \infty$, the combination of (5.56) and the Lebesgue differentiation theorem [SS05, Cor. 1.7 in Chap. 3] implies $\|A^*\vartheta\|_{L^2(\Omega)} = 0$. This, the vanishing trace $\gamma_{A^*}^{T}\vartheta|_T = 0$, the integration by parts formula (5.34), and the assumption $\mathbb{P}_0(T; \mathbb{R}^{m_A \times n_A}) \subset A(H(A, T))$ yield

$$\int_T \vartheta \,\mathrm{d}x = 0 \qquad \text{for all } \ell \in \mathbb{N} \text{ and } T \in \mathcal{T}_\ell.$$

The Lebesgue differentiation theorem proves $\|\vartheta\|_{L^2(\Omega)} = 0$ and concludes the proof. □

Theorem 5.1.41 (Convergence of $\mathcal{E}_{\rho,\ell}$)**.** *Suppose Assumption 5.1.21–5.1.22 and (5.54). Then, for all functions* $(v, \tau, \vartheta) \in Z := V \times W$ *and weights* $\rho > 0$,

$$\lim_{\ell\to\infty} \|\mathcal{E}_{\rho,\ell}(v, \tau, \gamma_{A^*}^{\mathcal{T}_\ell}\vartheta) - \vartheta\|_{H(A^*,\Omega,\rho)} = 0.$$

Proof. Let the function $(v, \tau, \vartheta) \in Z$ and the weight $\rho > 0$.
Step 1 (Boundedness of $\|b_\ell(v, \tau, \gamma_{A^*}^{\mathcal{T}_\ell}\vartheta; \bullet)\|_{Y_\rho(\mathcal{T}_\ell)^*}$*).* For all $\ell \in \mathbb{N}$ and $w^{\mathrm{pw}} \in H(A, \mathcal{T}_\ell)$, Corollary 5.1.32 and the Cauchy-Schwarz inequality imply

$$\begin{aligned}
&b_\ell(v, \tau, \gamma_{A^*}^{\mathcal{T}_\ell}\vartheta; w^{\mathrm{pw}}) = (Bv, A_{NC}w^{\mathrm{pw}})_{L^2(\Omega)} + (C\tau, w^{\mathrm{pw}})_{L^2(\Omega)} - \langle \gamma_{A^*}^{\mathcal{T}_\ell}\vartheta, w^{\mathrm{pw}}\rangle_{\partial\mathcal{T}_\ell} \\
&\le \|Bv\|_{L^2(\Omega)}\|A_{NC}w^{\mathrm{pw}}\|_{L^2(\Omega)} + \|C\tau\|_{L^2(\Omega)}\|w^{\mathrm{pw}}\|_{L^2(\Omega)} + \|\gamma_{A^*}^{\mathcal{T}_\ell}\vartheta\|_{\Gamma_{A^*,\rho}(\partial\mathcal{T}_\ell)}\|w^{\mathrm{pw}}\|_{Y_\rho(\mathcal{T}_\ell)} \\
&\le \Big((\rho^{-1}\|Bv\|^2_{L^2(\Omega)} + \|C\tau\|^2_{L^2(\Omega)})^{1/2} + \rho^{-1/2}\|\vartheta\|_{H(A,\Omega,\rho)}\Big)\|w^{\mathrm{pw}}\|_{Y_\rho(\mathcal{T}_\ell)}.
\end{aligned}$$

Step 2 (Orthogonality of the trial-to-test operators). Let $k \le \ell$, then (5.30a) and (5.53) imply $H(A, \mathcal{T}_k) \subset H(A, \mathcal{T}_\ell)$. This and the integration by parts formula (5.34) result in

$$\begin{aligned}
&(T_{\rho,k}(v, \tau, \gamma_{A^*}^{\mathcal{T}_k}\vartheta) - T_{\rho,\ell}(v, \tau, \gamma_{A^*}^{\mathcal{T}_\ell}\vartheta), T_{\rho,k}(v, \tau, \gamma_{A^*}^{\mathcal{T}_k}\vartheta))_{Y_\rho(\mathcal{T}_\ell)} \\
&\quad = b_k(v, \tau, \gamma_{A^*}^{\mathcal{T}_k}\vartheta; T_{\rho,k}(v, \tau, \gamma_{A^*}^{\mathcal{T}_k}\vartheta)) - b_\ell(v, \tau, \gamma_{A^*}^{\mathcal{T}_\ell}\vartheta; T_{\rho,k}(v, \tau, \gamma_{A^*}^{\mathcal{T}_k}\vartheta)) \\
&\quad = \langle \gamma_{A^*}^{\mathcal{T}_\ell}\vartheta, T_{\rho,k}(v, \tau, \gamma_{A^*}^{\mathcal{T}_k}\vartheta)\rangle_{\partial\mathcal{T}_\ell} - \langle \gamma_{A^*}^{\mathcal{T}_k}\vartheta, T_{\rho,k}(v, \tau, \gamma_{A^*}^{\mathcal{T}_k}\vartheta)\rangle_{\partial\mathcal{T}_k} = 0.
\end{aligned}$$

Step 3 (Convergence of $\mathcal{E}_{\rho,\ell}(v, \tau, \gamma_{A^*}^{\mathcal{T}_\ell}\vartheta)$*).* Let $k \le \ell$. The orthogonality from Step 2, the Pythagorean theorem, the definition of $\mathcal{E}_{\rho,\ell}$ in (5.47), and Theorem 5.1.31*(i)* show

$$\begin{aligned}
0 &\le \|\mathcal{E}_{\rho,\ell}(v, \tau, \gamma_{A^*}^{\mathcal{T}_\ell}\vartheta) - \mathcal{E}_{\rho,k}(v, \tau, \gamma_{A^*}^{\mathcal{T}_k}\vartheta)\|^2_{H(A^*,\Omega,\rho)} = \|T_{\rho,\ell}(v, \tau, \gamma_{A^*}^{\mathcal{T}_\ell}\vartheta) - T_{\rho,k}(v, \tau, \gamma_{A^*}^{\mathcal{T}_k}\vartheta)\|^2_{Y_\rho(\mathcal{T}_\ell)} \\
&= \|T_{\rho,\ell}(v, \tau, \gamma_{A^*}^{\mathcal{T}_\ell}\vartheta)\|^2_{Y_\rho(\mathcal{T}_\ell)} - \|T_{\rho,k}(v, \tau, \gamma_{A^*}^{\mathcal{T}_k}\vartheta)\|^2_{Y_\rho(\mathcal{T}_k)}. \qquad (5.58)
\end{aligned}$$

Hence, $(\|T_{\rho,\ell}(v, \tau, \gamma_{A^*}^{\mathcal{T}_\ell}\vartheta)\|^2_{Y_\rho(\mathcal{T}_\ell)})_{\ell\in\mathbb{N}}$ is a monotonically increasing and bounded (Step 1) sequence. Thus, the sequence converges. The convergence and (5.58) show that the sequence

$(\mathcal{E}_{\rho,\ell}(v,\sigma,\gamma_{A^*}^{\mathcal{T}_\ell}\vartheta))_{\ell\in\mathbb{N}}$ is a Cauchy sequence in the Hilbert space $H(A^*,\Omega)$. This implies the existence of a function $\Theta\in H(A^*,\Omega)$ with

$$\lim_{\ell\to\infty}\|\mathcal{E}_{\rho,\ell}(v,\tau,\gamma_{A^*}^{\mathcal{T}_\ell}\vartheta)-\Theta\|_{H(A^*,\Omega,\rho)}=0.$$

Step 4 ($\vartheta=\Theta$). Theorem 5.1.31*(ii)*, Corollary 5.1.32, and Step 3 prove

$$\begin{aligned}\|\gamma_{A^*}^{\mathcal{T}_\ell}(\vartheta-\Theta)\|_{\Gamma_{A^*,\rho}(\partial\mathcal{T}_\ell)}&=\|\gamma_{A^*}^{\mathcal{T}_\ell}(\mathcal{E}_{\rho,\ell}(v,\tau,\gamma_{A^*}^{\mathcal{T}_\ell}\vartheta)-\Theta)\|_{\Gamma_{A^*,\rho}(\partial\mathcal{T}_\ell)}\\&\le\|\mathcal{E}_{\rho,\ell}(v,\tau,\gamma_{A^*}^{\mathcal{T}_\ell}\vartheta)-\Theta\|_{H(A^*,\Omega,\rho)}\to0\quad\text{as}\quad\ell\to\infty.\end{aligned}\tag{5.59}$$

Furthermore, Corollary 5.1.32 shows that $\|\gamma_{A^*}^{\mathcal{T}_\ell}(\vartheta-\Theta)\|_{\Gamma_{A^*,\rho}(\partial\mathcal{T}_\ell)}$ is monotonically increasing in $\ell\in\mathbb{N}$. The combination of the monotonicity and (5.59) results in $\|\gamma_{A^*}^{\mathcal{T}_\ell}(\vartheta-\Theta)\|_{\Gamma_{A^*,\rho}(\partial\mathcal{T}_\ell)}=0$ for all $\ell\in\mathbb{N}$. Thus, (5.54) concludes the proof. □

Remark 5.1.42 (Convergent trace norm)**.** *Suppose Assumption 5.1.21–5.1.22 and (5.54), then Theorem 5.1.31(iii), Corollary 5.1.32, and Theorem 5.1.41 prove, for all functions $\vartheta\in W$ and weights $\rho>0$,*

$$\rho^{-1/2}\|\mathcal{E}_{\rho,\ell}(0,0,\gamma_{A^*}^{\mathcal{T}_\ell}\vartheta)\|_{H(A^*,\Omega,\rho)}=\|\gamma_{A^*}^{\mathcal{T}_\ell}\vartheta\|_{\Gamma_{A^*,\rho}(\partial\mathcal{T}_\ell)}\nearrow\rho^{-1/2}\|\vartheta\|_{H(A^*,\Omega,\rho)}\qquad\text{as }\ell\to\infty.$$

For all $(v,\tau,\vartheta)\in Z=V\times W$, $(w,\chi,t)\in X(\mathcal{T}_\ell)$, $\rho>0$, and $\ell\in\mathbb{N}$ define the norms

$$\|(v,\tau,\vartheta)\|_{Z_\rho}:=(\|(v,\tau)\|_V^2+\rho^{-1}\|\vartheta\|_{H(A^*,\Omega,\rho)}^2)^{1/2},\tag{5.60a}$$

$$\begin{aligned}\|(w,\chi,t)\|_{X_\rho(\mathcal{T}_\ell)}&:=(\|(w,\chi)\|_V^2+\rho^{-1}\|\mathcal{E}_{\rho,\ell}(w,\chi,t)\|_{H(A^*,\Omega,\rho)}^2)^{1/2}\\&=\|(w,\chi,\mathcal{E}_{\rho,\ell}(w,\chi,t))\|_{Z_\rho}.\end{aligned}\tag{5.60b}$$

Set for all $\ell\in\mathbb{N}$ the mesh-dependent constants

$$\begin{aligned}\beta_\rho(\mathcal{T}_\ell)&:=\inf_{(v,\tau,t)\in X(\mathcal{T}_\ell)\setminus\{0\}}\frac{\|b_\ell(v,\tau,t;\bullet)\|_{Y_\rho(\mathcal{T}_\ell)^*}}{\|(v,\tau,t)\|_{X_\rho(\mathcal{T}_\ell)}}\\&\le\sup_{(v,\tau,t)\in X(\mathcal{T}_\ell)\setminus\{0\}}\frac{\|b_\ell(v,\tau,t;\bullet)\|_{Y_\rho(\mathcal{T}_\ell)^*}}{\|(v,\tau,t)\|_{X_\rho(\mathcal{T}_\ell)}}=:\|b(\mathcal{T}_\ell)\|_\rho.\end{aligned}$$

Theorem 5.1.43 (Sharp estimate in (5.51))**.** *Suppose Assumption 5.1.21–5.1.22 and (5.54). The estimate in (5.51) is optimal in the sense that the constants*

$$\beta_\rho(\mathcal{T}_\ell)\searrow\beta_\rho\quad\text{and}\quad\|b(\mathcal{T}_\ell)\|_\rho\nearrow\|b\|_\rho\qquad\text{as }\ell\to\infty.$$

Proof. Let the weight $\rho>0$. For all $\varepsilon>0$ exists a function $0\neq(v,\tau,\vartheta)\in Z:=V\times W$ with $\|(v,\tau,\vartheta)\|_{a_\rho}/\|(v,\tau,\vartheta)\|_{Z_\rho}=\beta_\rho+\varepsilon$. Theorem 5.1.41, Theorem 5.1.31*(vi)*, and the triangle inequality prove

$$\lim_{\ell\to\infty}\frac{\|b_\ell(v,\tau,\gamma_{A^*}^{\mathcal{T}_\ell}\vartheta;\bullet)\|_{Y_\rho(\mathcal{T}_\ell)^*}}{\|(v,\tau,\gamma_{A^*}^{\mathcal{T}_\ell}\vartheta)\|_{X_\rho(\mathcal{T}_\ell)}}=\lim_{\ell\to\infty}\frac{\|(v,\tau,\mathcal{E}_{\rho,\ell}(v,\tau,\gamma_{A^*}^{\mathcal{T}_\ell}\vartheta)\|_{a_\rho}}{\|(v,\tau,\mathcal{E}_{\rho,\ell}(v,\tau,\gamma_{A^*}^{\mathcal{T}_\ell}\vartheta)\|_{Z_\rho}}=\frac{\|(v,\tau,\vartheta)\|_{a_\rho}}{\|(v,\tau,\vartheta)\|_{Z_\rho}}=\beta_\rho+\varepsilon.$$

This proves $\lim_{\ell\to\infty}\beta_\rho(\mathcal{T}_\ell)\le\beta_\rho$. The combination with $\beta_\rho\le\beta_\rho(\mathcal{T}_\ell)$ from (5.51a) for all $\ell\in\mathbb{N}$ results in $\lim_{\ell\to\infty}\beta(\mathcal{T}_\ell)=\beta_\rho$. Similar arguments prove $\lim_{\ell\to\infty}\|b(\mathcal{T}_\ell)\|_\rho=\|b\|_\rho$. □

Remark 5.1.44 (Product norm in X)**.** *In most publications, the norm in X reads*

$$\|(v,\tau,t)\|_X = (\|(v,\tau)\|_V^2 + \|t\|^2_{\Gamma_{A^*,1}(\partial\mathcal{T})})^{1/2} \qquad \text{for all } (v,\tau,t)\in X$$

with the minimal extension norm $\|t\|_{\Gamma_{A^*,1}(\partial\mathcal{T})} = \|\mathcal{E}_1(0,0,t)\|_{H(A^*,\Omega)}$ from Corollary 5.1.32. Theorem 5.1.31(iv)–(v) show, for all $(v,\tau,t)\in X$,

$$\begin{aligned}\|(v,\tau,t)\|_X^2 &\le \|(v,\tau)\|_V^2 + \|t\|^2_{\Gamma_{A^*,1}(\partial\mathcal{T})} + \|\mathcal{E}_1(v,\tau,0)\|^2_{H(A^*,\Omega)} = \|(v,\tau,t)\|^2_{X_1}\\ &\le \|(v,\tau)\|_V^2 + \|t\|^2_{\Gamma_{A^*,1}(\partial\mathcal{T})} + \|Bv\|^2_{L^2(\Omega)} + \|C\tau\|^2_{L^2(\Omega)}.\end{aligned}$$

If there exists a constant $\Lambda>0$ with $\|Bv\|^2_{L^2(\Omega)} + \|C\tau\|^2_{L^2(\Omega)} \le \Lambda\|(v,\tau)\|_V^2$ for all $(v,\tau)\in V$, the norms $\|\bullet\|_X$ and $\|\bullet\|_{X_1}$ are equivalent. Moreover, Theorem 5.1.41 implies that $\lim_{\ell\to\infty}\|\mathcal{E}_{1,\ell}(v,\tau,0)\|_{H(A^*,\Omega)} = 0$ and so the arguments from Theorem 5.1.43 result in

$$\inf_{(v,\tau,t)\in X(\mathcal{T}_\ell)\setminus\{0\}} \frac{\|b_\ell(v,\tau,t;\bullet)\|_{Y_1(\mathcal{T}_\ell)^*}}{(\|(v,\tau)\|_V^2 + \|t\|^2_{\Gamma_{A^*,1}(\partial\mathcal{T}_\ell)})^{1/2}} \searrow \beta_1 \qquad \text{as } \ell\to\infty.$$

Recall the norm $\|\bullet\|_{Z_\rho}$ in $Z = V\times W$ from (5.60a) with weight $\rho>0$ and let $V_h(\mathcal{T}_\ell)\subset V$ and $W_h(\mathcal{T}_\ell)\subset W$ be discrete subspaces with the density property

$$\lim_{\ell\to\infty}\min_{z_h\in V_h(\mathcal{T}_\ell)\times W_h(\mathcal{T}_\ell)}\|z-z_h\|_{Z_\rho} = 0 \qquad \text{for all } z\in Z. \tag{5.61}$$

For all $\ell\in\mathbb{N}$ and $\rho>0$ define the subspace

$$Z_{\rho,h}(\mathcal{T}_\ell) := \{(v_h,\tau_h,\mathcal{E}_{\rho,\ell}(v_h,\tau_h,\gamma^{\mathcal{T}_\ell}_{A^*}\vartheta_h)) \mid (v_h,\tau_h)\in V_h(\mathcal{T}_\ell) \text{ and } \vartheta_h\in W_h(\mathcal{T}_\ell)\}\subset Z.$$

Theorem 5.1.45 (Density of $Z_{\rho,h}(\mathcal{T}_\ell)$ in Z)**.** *Let the weight $\rho>0$ and suppose the assumptions of Theorem 5.1.41. Moreover, assume the density property (5.61) and let there exists a constant $\Lambda>0$ with*

$$\|Bv\|^2_{L^2(\Omega)} + \rho\|C\tau\|^2_{L^2(\Omega)} \le \Lambda\|(v,\tau)\|_V^2 \qquad \text{for all } (v,\tau)\in V. \tag{5.62}$$

Then

$$\lim_{\ell\to\infty}\min_{z_h\in Z_{\rho,h}(\mathcal{T}_\ell)}\|z-z_h\|_{Z_\rho} = 0 \qquad \text{for all } z\in Z.$$

Proof. Given $z=(v,\tau,\vartheta)\in Z$ and $\rho>0$, let $(v_{h,\ell},\tau_{h,\ell})\in V_h(\mathcal{T}_\ell)$ and $\vartheta_{h,\ell}\in W_h(\mathcal{T}_\ell)$ with $\|(v_{h,\ell},\tau_{h,\ell},\vartheta_{h,\ell})-(v,\tau,\vartheta)\|_{Z_\rho}\to 0$ as $\ell\to\infty$. The triangle inequality implies, for all $z_{h,\ell} = (v_{h,\ell},\tau_{h,\ell},\mathcal{E}_{\rho,\ell}(v_{h,\ell},\tau_{h,\ell},\gamma^{\mathcal{T}_\ell}_{A^*}\vartheta_{h,\ell}))\in Z_{h,\rho}(\mathcal{T}_\ell)$ and $\ell\in\mathbb{N}$, that

$$\begin{aligned}\|z-z_{h,\ell}\|^2_{Z_\rho} &= \|(v,\tau)-(v_{h,\ell},\tau_{h,\ell})\|_V^2 + \|\vartheta-\mathcal{E}_{\rho,\ell}(v_{h,\ell},\tau_{h,\ell},\gamma^{\mathcal{T}_\ell}_{A^*}\vartheta_{h,\ell})\|^2_{H(A^*,\Omega,\rho)}\\ &\le \|(v,\tau)-(v_{h,\ell},\tau_{h,\ell})\|_V^2 + (\|\mathcal{E}_{\rho,\ell}(v,\tau,0)\|_{H(A^*,\Omega,\rho)} + \|\mathcal{E}_{\rho,\ell}(v-v_{h,\ell},\tau-\tau_{h,\ell},0)\|_{H(A^*,\Omega,\rho)}\\ &\quad + \|\vartheta-\mathcal{E}_{\rho,\ell}(0,0,\gamma^{\mathcal{T}_\ell}_{A^*}\vartheta)\|_{H(A^*,\Omega,\rho)} + \|\mathcal{E}_{\rho,\ell}(0,0,\gamma^{\mathcal{T}_\ell}_{A^*}(\vartheta-\vartheta_{h,\ell}))\|_{H(A^*,\Omega,\rho)})^2.\end{aligned}$$

The convergence $\|(v_{h,\ell},\tau_{h,\ell},\vartheta_{h,\ell})-(v,\tau,\vartheta)\|_{Z_\rho}\to 0$, the combination of (5.62) and Theorem 5.1.31*(iv)*, and Theorem 5.1.41 prove that all summands tend to zero as $\ell\to\infty$. □

Paper	Problem	A^*	A
[CGHW14, Sec. 2.3]	Poisson	$\begin{pmatrix} 0 & -\operatorname{div} \\ -\nabla & \mathrm{id} \end{pmatrix}$	$\begin{pmatrix} 0 & \operatorname{div} \\ \nabla & \mathrm{id} \end{pmatrix}$
[GMO14, Eq. 1.1]	Helmholtz	$\begin{pmatrix} i\omega & \operatorname{div} \\ \nabla & i\omega \end{pmatrix}$	$\begin{pmatrix} i\omega & \operatorname{div} \\ \nabla & i\omega \end{pmatrix}$
[CDG16, Sec. 6.2]	Maxwell	$\begin{pmatrix} i\omega & -\operatorname{curl} \\ \operatorname{curl} & i\omega \end{pmatrix}$	$\begin{pmatrix} -i\omega & \operatorname{curl} \\ -\operatorname{curl} & -i\omega \end{pmatrix}$
[RBTD14, Sec. 2.2]	Stokes	$\begin{pmatrix} 0 & \nabla & -\operatorname{div} \\ \operatorname{div} & 0 & 0 \\ -\nabla & 0 & \mathrm{id} \end{pmatrix}$	$\begin{pmatrix} 0 & -\nabla & \operatorname{div} \\ -\operatorname{div} & 0 & 0 \\ \nabla & 0 & \mathrm{id} \end{pmatrix}$
[EW19, Sec. 2] [GS17, Sec. 2.1]	Wave	$\begin{pmatrix} \partial_t & \operatorname{div} \\ \nabla & \partial_t \end{pmatrix}$	$\begin{pmatrix} -\partial_t & -\operatorname{div} \\ -\nabla & -\partial_t \end{pmatrix}$

Table 5.2: Operators for existing ultra-weak DPG formulations (with frequency $\omega > 0$ and imaginary unit $i = \sqrt{-1}$)

5.1.6 Ultra-weak DPG

A special case of the abstract problem (5.29) seeks a solution u in a closed subspace $H_0(A^*, \Omega) \subset W \subset H(A^*, \Omega)$ to the problem

$$A^* u = f. \tag{5.63}$$

This problem occurs in ultra-weak DPG formulations, as for example in [CGHW14] for the Poisson model problem, in [GMO14] for the Helmholtz equation, in [CDG16] for the Maxwell equations, in [RBTD14] for the Stokes problem, and in [EW19, GS17] for the wave equation (cf. Table 5.2). The large scientific interest in ultra-weak DPG formulations motivates a more detailed analysis of this special case. The analysis focuses on the computation of the inf-sup constant β from (5.4) and the stability of the DPG method with respect to the weight $\rho > 0$ in the test norm $\|\bullet\|_{Y_\rho}$ from (5.45).

Recall the definitions from (5.39)–(5.42) and suppose Assumption 5.1.21–5.1.22. Moreover, let there exist a coercivity constant $c > 0$ with

$$c\|\tau\|^2_{L^2(\Omega)} \leq \|A^*\tau\|^2_{L^2(\Omega)} \qquad \text{for all } \tau \in W. \tag{5.64}$$

The bilinear form $b(\bullet, \bullet)$ from (5.42) reads, for all $(v, t) \in X := V \times \gamma^{\mathcal{T}}_{A^*} W$ with $V := L^2(\Omega; \mathbb{R}^{m_{A^*} \times n_{A^*}})$ and $w^{\mathrm{pw}} \in Y = H(A, \mathcal{T})$,

$$b(v, t; w^{\mathrm{pw}}) := (v, A_{NC} w^{\mathrm{pw}})_{L^2(\Omega)} - \langle t, w^{\mathrm{pw}} \rangle_{\partial\mathcal{T}}.$$

Theorem 5.1.29 proves the equivalence of (5.63) and the problem: Seek $(u, s) \in X$ with

$$b(u, s; w^{\mathrm{pw}}) = (f, w^{\mathrm{pw}})_{L^2(\Omega)} \qquad \text{for all } w^{\mathrm{pw}} \in H(A, \mathcal{T}). \tag{5.65}$$

For all $(v, t) \in X$, $w^{\mathrm{pw}} \in H(A, \mathcal{T})$, and weights $\rho > 0$ recall the operator $\mathcal{E}_\rho(v, t) := \mathcal{E}_\rho(v, 0, t)$ from Theorem 5.1.31 and the ρ-dependent norms

$$\|(v, t)\|_{X_\rho} := (\|v\|^2_{L^2(\Omega)} + \|A^* \mathcal{E}_\rho(v, t)\|^2_{L^2(\Omega)})^{1/2}, \tag{5.66a}$$

$$\|w^{\mathrm{pw}}\|_{Y_\rho} := (\|w^{\mathrm{pw}}\|^2_{L^2(\Omega)} + \rho\|A_{NC} w^{\mathrm{pw}}\|^2_{L^2(\Omega)})^{1/2}. \tag{5.66b}$$

Given the constant $0 < c$ from (5.64) and a weight $0 < \rho$, set the constants

$$\beta_\rho^2 := \frac{1 + c + \rho c - \sqrt{(1 + c + \rho c)^2 - 4\rho c^2}}{2\rho c}, \tag{5.67a}$$

$$\|b\|_\rho^2 := \frac{1 + c + \rho c + \sqrt{(1 + c + \rho c)^2 - 4\rho c^2}}{2\rho c}. \tag{5.67b}$$

Theorem 5.1.46 (Stability constants)**.** *Suppose Assumption 5.1.21–5.1.22 and (5.64). For all weights $\rho > 0$ the constants β_ρ and $\|b\|_\rho$ from (5.67) satisfy*

$$\begin{aligned} 0 < \beta_\rho \le \beta_\rho(\mathcal{T}) &:= \inf_{x \in X\setminus\{0\}} \sup_{y \in Y\setminus\{0\}} \frac{b(x,y)}{\|x\|_{X_\rho}\|y\|_{Y_\rho}} \\ &\le \sup_{x \in X\setminus\{0\}} \sup_{y \in Y\setminus\{0\}} \frac{b(x,y)}{\|x\|_{X_\rho}\|y\|_{Y_\rho}} := \|b(\mathcal{T})\|_\rho \le \|b\|_\rho < \infty. \end{aligned} \tag{5.68}$$

Proof. Step 1 (Calculations). Given a constant $\gamma > 0$ and a weight $\rho > 0$, set the function

$$0 < g(\alpha) := \frac{\rho^{-1}(1-\alpha)^2 + \gamma}{\alpha^2 + \gamma} \qquad \text{for all } \alpha \in \mathbb{R}.$$

The derivative of g reads

$$g'(\alpha) = 2\,\frac{(\alpha - 1)(\alpha + \gamma) - \alpha\gamma\rho}{(\alpha^2 + \gamma)^2\rho} \qquad \text{for all } \alpha > 0.$$

The roots of g' lead the extrema

$$g\left(\frac{1 + \rho\gamma - \gamma}{2} + \sqrt{\frac{(1 + \rho\gamma - \gamma)^2}{4} + \gamma}\right) = \frac{1 + \gamma + \rho\gamma - \sqrt{(1 + \gamma + \rho\gamma)^2 - 4\rho\gamma^2}}{2\rho\gamma}, \tag{5.69a}$$

$$g\left(\frac{1 + \rho\gamma - \gamma}{2} - \sqrt{\frac{(1 + \rho\gamma - \gamma)^2}{4} + \gamma}\right) = \frac{1 + \gamma + \rho\gamma + \sqrt{(1 + \gamma + \rho\gamma)^2 - 4\rho\gamma^2}}{2\rho\gamma}. \tag{5.69b}$$

Since $g(\alpha)$ tends to ρ^{-1} as $|\alpha| \to \infty$ and has exactly two extrema, (5.69a) is a global minimum and (5.69b) is a global maximum. The minimum (5.69a) increases monotonically in $\gamma > 0$ and the maximum (5.69b) decreases monotonically in $\gamma > 0$. Moreover, $g(\alpha)$ tends to ρ^{-1} as $|\alpha| \to \infty$ and (5.69a) is a global minimum imply

$$\frac{1 + \gamma + \rho\gamma - \sqrt{(1 + \gamma + \rho\gamma)^2 - 4\rho\gamma^2}}{2\rho\gamma} < \frac{1}{\rho}. \tag{5.70}$$

Step 2 (Proof of (5.68)). Let the weight $\rho > 0$ and set $Z := V \times W$ with squared norm $\|(v,\tau)\|_Z^2 := \|v\|_{L^2(\Omega)}^2 + \|A^*\tau\|_{L^2(\Omega)}^2$ for all $(v,\tau) \in Z$. Theorem 5.1.31*(vi)* and the definition of $\|\bullet\|_{X_\rho}$ in (5.66a) imply

$$\frac{\|b(w,t;\bullet)\|_{Y_\rho^*}^2}{\|(w,t)\|_{X_\rho}^2} = \frac{\rho^{-1}\|w - \mathcal{E}_\rho(w,t)\|_{L^2(\Omega)}^2 + \|A^*\mathcal{E}_\rho(w,t)\|_{L^2(\Omega)}^2}{\|(w, \mathcal{E}_\rho(w,t))\|_Z^2} \quad \text{for all } (w,t) \in X. \tag{5.71}$$

Since $\mathcal{E}_\rho(w,t) \in W$ for all $(w,t) \in X$, the identity in (5.71) proves

$$\begin{aligned}\inf_{(v,\tau)\in Z\setminus\{0\}} \frac{\rho^{-1}\|v-\tau\|^2_{L^2(\Omega)} + \|A^*\tau\|^2_{L^2(\Omega)}}{\|(v,\tau)\|^2_Z} &\le \frac{\|b(w,t;\bullet)\|^2_{Y_\rho^*}}{\|(w,t)\|^2_{X_\rho}} \\ &\le \sup_{(v,\tau)\in Z\setminus\{0\}} \frac{\rho^{-1}\|v-\tau\|^2_{L^2(\Omega)} + \|A^*\tau\|^2_{L^2(\Omega)}}{\|(v,\tau)\|^2_Z}.\end{aligned} \tag{5.72}$$

Let $\tau \in W \subset V = L^2(\Omega;\mathbb{R}^{m_{A^*}\times n_{A^*}})$ with $\|\tau\|_{L^2(\Omega)} = 1$. Any function $v \in V$ decomposes into $v = \alpha\tau + w$ with $\alpha \in \mathbb{R}$ and orthogonal function $w \in V(\tau) := \{\xi \in V \mid (\tau,\xi)_{L^2(\Omega)} = 0\}$. The Pythagorean theorem results in

$$\inf_{v\in V} \frac{\rho^{-1}\|v-\tau\|^2_{L^2(\Omega)} + \|A^*\tau\|^2_{L^2(\Omega)}}{\|(v,\tau)\|^2_Z} = \inf_{\alpha\in\mathbb{R}, w\in V(\tau)} \frac{\rho^{-1}((1-\alpha)^2 + \|w\|^2_{L^2(\Omega)}) + \|A^*\tau\|^2_{L^2(\Omega)}}{\alpha^2 + \|w\|^2_{L^2(\Omega)} + \|A^*\tau\|^2_{L^2(\Omega)}}.$$

The identity in (5.69a) with $\gamma := \|A^*\tau\|^2_{L^2(\Omega)}$ and (5.70) yield

$$\inf_{\alpha\in\mathbb{R}} \frac{\rho^{-1}(1-\alpha)^2 + \|A^*\tau\|^2_{L^2(\Omega)}}{\alpha^2 + \|A^*\tau\|^2_{L^2(\Omega)}} = \frac{1+\gamma+\rho\gamma - \sqrt{(1+\gamma+\rho\gamma)^2 - 4\rho\gamma^2}}{2\rho\gamma} < \rho^{-1}. \tag{5.73}$$

For all $\alpha \in \mathbb{R}$ with $(\rho^{-1}(1-\alpha)^2+\gamma)/(\alpha^2+\gamma) < \rho^{-1}$ the function $(\rho^{-1}(1-\alpha)^2 + \rho^{-1}\delta + \gamma)/(\alpha^2+\delta+\gamma)$ increases monotonically in $\delta \ge 0$. This and (5.73) imply

$$\frac{1+\gamma+\rho\gamma - \sqrt{(1+\gamma+\rho\gamma)^2 - 4\rho\gamma^2}}{2\rho\gamma} = \inf_{\alpha\in\mathbb{R}, w\in V(\tau)} \frac{\rho^{-1}((1-\alpha)^2 + \|w\|^2_{L^2(\Omega)}) + \|A^*\tau\|^2_{L^2(\Omega)}}{\alpha^2 + \|w\|^2_{L^2(\Omega)} + \|A^*\tau\|^2_{L^2(\Omega)}}.$$

This identity, $\inf_{v\in V\setminus\{0\}} \rho^{-1}\|v\|^2_{L^2(\Omega)}/\|(v,0)\|^2_{Z_\rho} = \rho^{-1}$, $c \le \gamma$ from (5.64), and the monotonicity of the minimum (5.70) with respect to γ prove

$$\frac{1+c+\rho c - \sqrt{(1+c+\rho c)^2 - 4\rho c^2}}{2\rho c} \le \inf_{(v,\tau)\in Z\setminus\{0\}} \frac{\rho^{-1}\|v-\tau\|^2_{L^2(\Omega)} + \|A^*\tau\|^2_{L^2(\Omega)}}{\|(v,\tau)\|^2_Z}. \tag{5.74}$$

Similar arguments result in

$$\sup_{(v,\tau)\in Z\setminus\{0\}} \frac{\rho^{-1}\|v-\tau\|^2_{L^2(\Omega)} + \|A^*\tau\|^2_{L^2(\Omega)}}{\|(v,\tau)\|^2_Z} \le \frac{1+c+\rho c + \sqrt{(1+c+\rho c)^2 - 4\rho c^2}}{2\rho c}. \tag{5.75}$$

The combination of (5.72) and (5.74)–(5.75) concludes the proof. □

Remark 5.1.47 (Sharp estimate)**.** *Let* $0 < c = \inf_{\tau\in W\setminus\{0\}} \|A^*\tau\|^2_{L^2(\Omega)}/\|\tau\|^2_{L^2(\Omega)}$*, then (5.74)–(5.75) holds with equality. If in addition the assumptions of Theorem 5.1.43 hold, Theorem 5.1.43 shows the convergence* $\beta_\rho(\mathcal{T}_\ell) \searrow \beta_\rho$ *and* $\|b(\mathcal{T}_\ell)\|_\rho \nearrow \|b\|_\rho$ *as* $\ell \to \infty$*.*

Theorem 5.1.48 (Uniqueness condition)**.** *Suppose Assumption 5.1.21–5.1.22 and (5.64), then the uniqueness condition* (**H1**) *holds, that is,*

$$\{x \in X \mid b(x, w^{\mathrm{pw}}) = 0 \text{ for all } w^{\mathrm{pw}} \in H(A,\mathcal{T})\} = \{0\}.$$

Proof. Let $x = (v, t) \in X$ with $b(x, w^{\mathrm{pw}}) = 0$ for all $w^{\mathrm{pw}} \in H(A, \mathcal{T})$. Lemma 5.1.30 implies

$$0 = b(x, w) = (v, A_{NC} w)_{L^2(\Omega)} + \langle t, w \rangle_{\partial\mathcal{T}} = (v, Aw)_{L^2(\Omega)} \quad \text{for all } w \in H_0(A, \Omega) \subset H(A, \mathcal{T}).$$

Thus, the characterization in (5.31a) shows $v \in H(A^*, \Omega)$ with $A^* v = 0$. This allows for a piecewise integration by parts (5.34) and so yields, for all $w^{\mathrm{pw}} \in H(A, \mathcal{T})$,

$$0 = b(x, w^{\mathrm{pw}}) = (v, A_{NC} w^{\mathrm{pw}})_{L^2(Q)} - \langle t, w^{\mathrm{pw}} \rangle_{\partial\mathcal{T}} = \langle \gamma_{A^*}^{\mathcal{T}} v - t, w^{\mathrm{pw}} \rangle_{\partial\mathcal{T}}. \tag{5.76}$$

Let $\vartheta \in W$ with trace $\gamma_{A^*}^{\mathcal{T}} \vartheta = t \in \gamma_{A^*}^{\mathcal{T}} W$. An integration by parts in (5.76) shows, for all $w \in H(A, \Omega) \subset H(A, \mathcal{T})$,

$$0 = \langle \gamma_{A^*}^{\mathcal{T}} v - t, w \rangle_{\partial\mathcal{T}} = (v - \vartheta, Aw)_{L^2(\Omega)} - (A^*(v - \vartheta), w)_{L^2(\Omega)}.$$

Thus, (5.38b) results in $v - \vartheta \in H_0(A^*, \Omega) \subset W$. The combination of $A^* v = 0$, $v = (v - \vartheta) + \vartheta \in W$, and (5.64) proves $v = 0$. This and (5.76) show $t = 0$. □

The combination of Theorem 5.1.46 and Theorem 5.1.48 results in the well-posedness of the variational problem (5.65). If the discrete spaces $X_h \subset X$ and $Y_h \subset H(A, \mathcal{T})$ allow for an operator $P : Y \to Y_h$ with (5.5), Theorem 5.1.1 leads to a priori error estimates with respect to the ρ-dependent norm $\|\bullet\|_{X_\rho}$ from (5.66a). The following theorem shows the equivalence of $\|\bullet\|_{X_\rho}$ and a ρ-independent norm.

Theorem 5.1.49 (ρ-independent norm $\|A^* \mathcal{E}_\infty(\bullet)\|_{L^2(\Omega)}$)**.** *Suppose the assumptions from Theorem 5.1.46. For all $(v, t) \in X$ the function $\mathcal{E}_\rho(v, t)$ converges towards a function $\mathcal{E}_\infty(t) \in W$ as $\rho \to \infty$, that is*

$$\lim_{\rho \to \infty} \|A^*(\mathcal{E}_\infty(t) - \mathcal{E}_\rho(v, t))\|_{L^2(\Omega)} = 0.$$

The function $\mathcal{E}_\infty(t) \in W$ satisfies, for all $(v, t) \in X$,

(i) $\gamma_{A^*}^{\mathcal{T}} \mathcal{E}_\infty(t) = t$,

(ii) $\|A^* \mathcal{E}_\infty(t)\|_{L^2(\Omega)} = \min\{\|A^* \vartheta\|_{L^2(\Omega)} \mid \vartheta \in W \text{ and } \gamma_{A^*}^{\mathcal{T}} \vartheta = t\}$,

(iii) $\|A^* \mathcal{E}_\infty(t)\|_{L^2(\Omega)}^2 \le \|A^* \mathcal{E}_\rho(v, t)\|_{L^2(\Omega)}^2 \le \rho^{-1} \|\mathcal{E}_\rho(v, t)\|_{H(A^*, \Omega, \rho)}^2$
$\le (1 + \rho^{-1} c^{-1}) \|A^* \mathcal{E}_\infty(t)\|_{L^2(\Omega)}^2 + \rho^{-1} \|v\|_{L^2(\Omega)}^2$,

(iv) $\|A^*(\mathcal{E}_\rho(v, t) - \mathcal{E}_\infty(t))\|_{L^2(\Omega)}^2 \le \rho^{-1}(c^{-1} \|A^* \mathcal{E}_\infty(t)\|_{L^2(\Omega)}^2 + \|v\|_{L^2(\Omega)}^2)$.

Proof. Let $(v, t) \in X$ and define the infimum

$$0 \le \gamma := \inf\{\|A^* \vartheta\|_{L^2(\Omega)} \mid \vartheta \in W \text{ and } \gamma_{A^*}^{\mathcal{T}} \vartheta = t\}.$$

Step 1 (Proof of (iii)). Let $(\vartheta_n)_{n \in \mathbb{N}} \subset \{\vartheta \in W \mid \gamma_{A^*}^{\mathcal{T}} \vartheta = t\}$ be an infimizing sequence in the sense that $\|A^* \vartheta_n\|_{L^2(\Omega)} \to \gamma$. Theorem 5.1.31 and (5.64) imply

$$\begin{aligned} \gamma^2 &\le \|A^* \mathcal{E}_\rho(v, t)\|_{L^2(\Omega)}^2 \le \|A^* \mathcal{E}_\rho(0, t)\|_{L^2(\Omega)}^2 + \rho^{-1} \|\mathcal{E}_\rho(0, t)\|_{L^2(\Omega)}^2 + \rho^{-1} \|v\|_{L^2(\Omega)}^2 \\ &\le \rho^{-1} \|\vartheta_n\|_{H(A^*, \Omega, \rho)}^2 + \rho^{-1} \|v\|_{L^2(\Omega)}^2 \\ &\le (1 + \rho^{-1} c^{-1}) \|A^* \vartheta_n\|_{L^2(\Omega)}^2 + \rho^{-1} \|v\|_{L^2(\Omega)}^2 \qquad \text{for all } n \in \mathbb{N} \text{ and } \rho > 0. \end{aligned}$$

Passing to the limit $n \to \infty$ proves

$$\gamma^2 \le \|A^* \mathcal{E}_\rho(v, t)\|_{L^2(\Omega)}^2 \le (1 + \rho^{-1} c^{-1}) \gamma^2 + \rho^{-1} \|v\|_{L^2(\Omega)}^2 \qquad \text{for all } \rho > 0.$$

Thus, $\lim_{\rho\to\infty}\|A^*\mathcal{E}_\rho(v,t)\|_{L^2(\Omega)} = \gamma$.

Step 2 ($\mathcal{E}_\rho(0,t) \to \mathcal{E}_\infty(t)$). Given $0 < \rho_1 < \rho_2$, Theorem 5.1.31*(ii)* and (5.48) show $(\mathcal{E}_{\rho_2}(0,t), \mathcal{E}_{\rho_2}(0,t) - \mathcal{E}_{\rho_1}(0,t))_{H(A^*,\Omega,\rho_2)} = 0$. This proves

$$\begin{aligned}&(A^*\mathcal{E}_{\rho_2}(0,t), A^*\mathcal{E}_{\rho_1}(0,t))_{L^2(\Omega)}\\ &= \|A^*\mathcal{E}_{\rho_2}(0,t)\|^2_{L^2(\Omega)} + \rho_2^{-1}\|\mathcal{E}_{\rho_2}(0,t)\|^2_{L^2(\Omega)} - \rho_2^{-1}(\mathcal{E}_{\rho_2}(0,t), \mathcal{E}_{\rho_1}(0,t))_{L^2(\Omega)}.\end{aligned}$$

This identity and the Cauchy-Schwarz inequality result in

$$\begin{aligned}&\|A^*(\mathcal{E}_{\rho_2}(0,t) - \mathcal{E}_{\rho_1}(0,t))\|^2_{L^2(\Omega)}\\ &= \|A^*\mathcal{E}_{\rho_2}(0,t)\|^2_{L^2(\Omega)} + \|A^*\mathcal{E}_{\rho_1}(0,t)\|^2_{L^2(\Omega)} - 2\,(A^*\mathcal{E}_{\rho_2}(0,t), A^*\mathcal{E}_{\rho_1}(0,t))_{L^2(\Omega)}\\ &= \|A^*\mathcal{E}_{\rho_1}(0,t)\|^2_{L^2(\Omega)} - \|A^*\mathcal{E}_{\rho_2}(0,t)\|^2_{L^2(\Omega)} - 2\rho_2^{-1}\|\mathcal{E}_{\rho_2}(0,t)\|^2_{L^2(\Omega)}\\ &\quad + 2\rho_2^{-1}(\mathcal{E}_{\rho_2}(0,t), \mathcal{E}_{\rho_1}(0,t))_{L^2(\Omega)}\\ &\le \|A^*\mathcal{E}_{\rho_1}(0,t)\|^2_{L^2(\Omega)} - \|A^*\mathcal{E}_{\rho_2}(0,t)\|^2_{L^2(\Omega)} + \rho_2^{-1}\|\mathcal{E}_{\rho_1}(0,t)\|^2_{L^2(\Omega)} - \rho_2^{-1}\|\mathcal{E}_{\rho_2}(0,t)\|^2_{L^2(\Omega)}.\end{aligned} \tag{5.77}$$

Since $\lim_{\rho\to\infty}\|A^*\mathcal{E}_\rho(0,t)\|^2_{L^2(\Omega)} = \gamma^2$ and $\lim_{\rho\to\infty}\rho^{-1}\|\mathcal{E}_\rho(0,t)\|^2_{L^2(\Omega)} = 0$, the inequality in (5.77) proves that $(\mathcal{E}_\rho(0,t))_{\rho>0}$ is a Cauchy sequence in the Hilbert space W. Thus, there exists an element $\mathcal{E}_\infty(t) \in W$ with

$$\lim_{\rho\to\infty}\|A^*(\mathcal{E}_\rho(0,t) - \mathcal{E}_\infty(t))\|_{L^2(\Omega)} = 0 = \lim_{\rho\to\infty}\rho^{-1}\|\mathcal{E}_\rho(0,t) - \mathcal{E}_\infty(t)\|_{H(A^*,\Omega,\rho)}. \tag{5.78}$$

The combination with $\lim_{\rho\to\infty}\|A^*\mathcal{E}_\rho(0,t)\|_{L^2(\Omega)} = \gamma$ from Step 1 proves $\gamma = \|A^*\mathcal{E}_\infty(t)\|_{L^2(\Omega)}$ and validates *(ii)*.

Step 3 ($\mathcal{E}_\rho(v,t) \to \mathcal{E}_\infty(t)$). Theorem 5.1.31*(iv)*, (5.78), and the triangle inequality prove

$$\lim_{\rho\to\infty}\|A^*(\mathcal{E}_\rho(v,t) - \mathcal{E}_\infty(t))\|_{L^2(\Omega)} \le \lim_{\rho\to\infty}\|A^*(\mathcal{E}_\rho(0,t) - \mathcal{E}_\infty(t))\|_{L^2(\Omega)} + \rho^{-1/2}\|v\|_{L^2(\Omega)} = 0.$$

Step 4 (Proof of (i)). Theorem 5.1.31*(ii)* and Corollary 5.1.32 prove

$$\begin{aligned}&\|\gamma^{\mathcal{T}}_{A^*}\mathcal{E}_\infty(t) - t\|_{\Gamma_{A^*,1}(\partial\mathcal{T})} = \|\gamma^{\mathcal{T}}_{A^*}(\mathcal{E}_\infty(t) - \mathcal{E}_\rho(0,t))\|_{\Gamma_{A^*,1}(\partial\mathcal{T})}\\ &\le \|\mathcal{E}_\infty(t) - \mathcal{E}_\rho(0,t)\|_{H(A^*,\Omega,1)} \le (1+c^{-1})\|A^*(\mathcal{E}_\infty(t) - \mathcal{E}_\rho(0,t))\|^2_{L^2(\Omega)} \to 0 \quad \text{as } \rho\to\infty.\end{aligned}$$

Step 5 (Proof of (iv)). Since $\mathcal{E}_\infty(t)$ satisfies *(ii)*, the characterization of best approximations (see Lemma 3.1.5), the identity in *(i)*, and Theorem 5.1.31*(ii)* prove

$$(A^*\mathcal{E}_\infty(t), A^*(\mathcal{E}_\infty(t) - \mathcal{E}_\rho(v,t)))_{L^2(\Omega)} = 0.$$

This orthogonality and *(iii)* result in

$$\begin{aligned}&\|A^*(\mathcal{E}_\infty(t) - \mathcal{E}_\rho(v,t))\|^2_{L^2(\Omega)} = \|A^*\mathcal{E}_\rho(v,t)\|^2_{L^2(\Omega)} - \|A^*\mathcal{E}_\infty(t)\|^2_{L^2(\Omega)}\\ &\le \rho^{-1}c^{-1}\|A^*\mathcal{E}_\infty(t)\|^2_{L^2(\Omega)} + \rho^{-1}\|v\|^2_{L^2(\Omega)}.\end{aligned}$$

□

Remark 5.1.50 (Choice of ρ)**.** *The a priori estimate from Theorem 5.1.1, the constants from Theorem 5.1.46, and the equivalence of norms from Theorem 5.1.49 suggest a weight that minimizes the constant*

$$C^2_{\mathrm{Stab}}(\rho) := \frac{\|b\|^2_\rho}{\beta^2_\rho} = \frac{1+c+\rho c+\sqrt{(1+c+\rho c)^2-4\rho c^2}}{1+c+\rho c-\sqrt{(1+c+\rho c)^2-4\rho c^2}} \qquad \textit{over all } \rho>0$$

in the a priori estimate from Theorem 5.1.1. The minimizer reads $\rho_{\min} = (c+1)/c$ and $C^2_{\mathrm{Stab}}(\rho_{\min}) = (2+c+2\sqrt{1+c})/c$.

Let $(u,s) \in X$ denote the exact solution to the variational problem (5.65). Let $X_h = V_h \times \Gamma_{A^*,h}(\partial\mathcal{T}) \subset X$ be a discrete subspace and define the solutions $(u_{h,\rho}, s_{h,\rho}) \in X_h$ to the idealized DPG method with weight ρ, that is

$$(u_{h,\rho}, s_{h,\rho}) = \operatorname*{arg\,min}_{x_h \in X_h} \|b(u,s;\bullet) - b(x_h,\bullet)\|_{Y_\rho^*} \qquad \text{for all } \rho > 0. \tag{5.79}$$

The discussion in [GMO14, Sec. 3.3] states the asymptotic best approximation result for an ultra-weak formulation of the Helmholtz equation (with $\lim'_{\rho\to\infty} := \lim'_{\rho^{-1}\to 0}$ from Definition 3.1.6)

$$\lim_{\rho\to\infty}{}' \frac{\|A^*\mathcal{E}_\infty(s - s_{h,\rho})\|_{L^2(\Omega)}}{\min_{t_h \in \Gamma_{A^*,h}(\partial\mathcal{T})} \|A^*\mathcal{E}_\infty(s - t_h)\|_{L^2(\Omega)}} = 1.$$

The remainder of this section generalizes this result.

Lemma 5.1.51 (Bounds for $\|b(v,t;\bullet)\|_{Y_\rho^*}$)**.** *Suppose the assumptions from Theorem 5.1.46 and recall the operator $\mathcal{E}_\infty$ from Theorem 5.1.49. For all $(v,t) \in X$ and weights $\rho > 0$*

$$\|A^*\mathcal{E}_\infty(t)\|_{L^2(\Omega)} \le \|b(v,t;\bullet)\|_{Y_\rho^*} \le (1 + \rho^{-1}c^{-1})^{1/2}\|A^*\mathcal{E}_\infty(t)\|_{L^2(\Omega)} + \rho^{-1/2}\|v\|_{L^2(\Omega)}.$$

Proof. Step 1 (Upper bound). Let $(v,t) \in X$ and $\rho > 0$. Theorem 5.1.49*(iii)*, Theorem 5.1.31*(iv)–(vi)*, and the Cauchy-Schwarz inequality prove

$$\begin{aligned}
\|b(0,t;\bullet)\|_{Y_\rho^*}^2 &= \rho^{-1}\|\mathcal{E}_\rho(0,t)\|_{L^2(\Omega)}^2 + \|A^*\mathcal{E}_\rho(0,t)\|_{L^2(\Omega)}^2 = \rho^{-1}\|\mathcal{E}_\rho(0,t)\|_{H(A^*,\Omega,\rho)}^2 \\
&\le \rho^{-1}\|\mathcal{E}_\infty(t)\|_{H(A^*,\Omega,\rho)}^2 \le (1 + \rho^{-1}c^{-1})\|A^*\mathcal{E}_\infty(t)\|_{L^2(\Omega)}^2, \qquad (5.80)\\
\|b(v,0;\bullet)\|_{Y_\rho^*}^2 &= \|(v, A_{NC}\bullet)_{L^2(\Omega)}\|_{Y_\rho^*}^2 \le \sup_{w^{\mathrm{pw}} \in Y\setminus\{0\}} \frac{\|v\|_{L^2(\Omega)}^2 \|A_{NC}w^{\mathrm{pw}}\|_{L^2(\Omega)}^2}{\|w^{\mathrm{pw}}\|_{Y_\rho}^2} \le \rho^{-1}\|v\|_{L^2(\Omega)}^2.
\end{aligned}$$

The combination of these two inequalities and the triangle inequality $\|b(v,t;\bullet)\|_{Y_\rho^*} \le \|b(0,t;\bullet)\|_{Y_\rho^*} + \|b(v,0;\bullet)\|_{Y_\rho^*}$ yield the upper bound in Lemma 5.1.51.

Step 2 (Lower bound). Let $(v,t) \in X$ and $\rho > 0$. Theorem 5.1.31*(vi)*, Theorem 5.1.49*(iii)*, (5.80), and the reverse triangle inequality prove

$$\|A^*\mathcal{E}_\infty(t)\|_{L^2(\Omega)} - \rho^{-1/2}\|v\|_{L^2(\Omega)} \le \|b(0,t;\bullet)\|_{Y_\rho^*} - \|b(v,0;\bullet)\|_{Y_\rho^*} \le \|b(v,t;\bullet)\|_{Y_\rho^*}.$$

Since $\|b(v,t;\bullet)\|_{Y_\rho^*}$ decreases monotonically in ρ, it holds $\|A^*\mathcal{E}_\infty(t)\|_{L^2(\Omega)} - \rho'^{-1/2}\|v\|_{L^2(\Omega)} \le \|b(v,t;\bullet)\|_{Y_\rho^*}$ for all $\rho < \rho'$. Passing to the limit $\rho' \to \infty$ results in the lower bound. □

Theorem 5.1.52 (Best approximation in $\|A^*\mathcal{E}_\infty(\bullet)\|_{L^2(\Omega)}$)**.** *Suppose the assumptions from Theorem 5.1.46 and recall the operator $\mathcal{E}_\infty$ from Theorem 5.1.49. The error $\|A^*\mathcal{E}_\infty(s - s_{h,\rho})\|_{L^2(\Omega)}$ with $s_{h,\rho}$ from (5.79) converges to the best approximation error as $\rho \to \infty$. More precisely,*

$$\begin{aligned}
\|A^*\mathcal{E}_\infty(s - s_{h,\rho})\|_{L^2(\Omega)} \le (1 + \rho^{-1}c^{-1})^{1/2} &\min_{t_h \in \Gamma_{A^*,h}(\partial\mathcal{T})} \|A^*\mathcal{E}_\infty(s - t_h)\|_{L^2(\Omega)} \\
&+ \rho^{-1/2} \min_{v_h \in V_h} \|u - v_h\|_{L^2(\Omega)}.
\end{aligned} \tag{5.81}$$

Proof. The identity in (5.79) and Lemma 5.1.51 imply, for all $\rho > 0$,

$$\begin{aligned}
&\|A^*\mathcal{E}_\infty(s - s_{h,\rho})\|_{L^2(\Omega)} \le \|b(u,s;\bullet) - b(u_{h,\rho}, s_{h,\rho};\bullet)\|_{Y_\rho^*} = \min_{x_h \in X_h} \|b(u,s;\bullet) - b(x_h,\bullet)\|_{Y_\rho^*} \\
&\le (1 + \rho^{-1}c^{-1})^{1/2} \min_{t_h \in \Gamma_{A^*,h}(\partial\mathcal{T})} \|A^*\mathcal{E}_\infty(s - t_h)\|_{L^2(\Omega)} + \rho^{-1/2} \min_{v_h \in V_h} \|u - v_h\|_{L^2(\Omega)}.
\end{aligned}$$

□

5.2 Application of DPG

5.2.1 Asymptotic exact DPG

This section utilizes the relation of the DPG method and the LSFEM from Theorem 5.1.35 to apply the asymptotic exactness results from Section 3.1.1 to a primal DPG formulation for the Poisson model problem (2.3) and Helmholtz equation (2.7). The problem reads: Given a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ with $d \in \mathbb{N}$, a frequency $\omega \geq 0$ ($\omega = 0$ for the Poisson model problem and $\omega > 0$ for the Helmholtz equation), and a right-hand side $f \in L^2(\Omega)$, seek the weak solution $u \in H^1_0(\Omega)$ to

$$-\Delta u - \omega^2 u = f. \tag{5.82}$$

Set the operators $A := B := \nabla$, $A^* := -\mathrm{div}$, $C := -\omega^2$ and let $u = \sigma$. Then (5.82) equals the abstract problem (5.29), which reads

$$A^* B u + C\sigma = f. \tag{5.83}$$

Theorem 5.1.15 and the definition of the (weak) differential operators ∇ and div in Section 2.1 imply Assumption 5.1.21–5.1.22 and so the design in Section 5.1.4 applies. Given a partition $\mathcal{T}$ of the bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ into a finite number of disjoint, non-empty, and convex Lipschitz domains with (5.25) (for example a regular triangulation of Ω into triangles or tetrahedra, see Remark 5.1.18), the design leads to the space $Y = H(A, \mathcal{T}) = H^1(\mathcal{T}) := \{w^{\mathrm{pw}} \in L^2(\Omega) \mid w^{\mathrm{pw}}|_T \in H^1(T) \text{ for all } T \in \mathcal{T}\}$ with norm $\|\bullet\|_Y^2 = \|\bullet\|_{L^2(\Omega)}^2 + \|\nabla_{NC}\bullet\|_{L^2(\Omega)}^2$ and the space

$$\mathcal{X} := H(B, \Omega) \times H(C, \Omega) \times \Gamma_{A^*}(\partial\mathcal{T}) = H^1(\Omega) \times L^2(\Omega) \times H^{-1/2}(\partial\mathcal{T})$$

with trace space $H^{-1/2}(\partial\mathcal{T}) = \gamma_\nu^{\mathcal{T}} H(\mathrm{div}, \Omega)$ from page 67. Moreover, it results in the bilinear forms $\langle \bullet, \bullet \rangle_{\partial\mathcal{T}} : H^{-1/2}(\partial\mathcal{T}) \times Y \to \mathbb{R}$ from (5.42a) and $b : \mathcal{X} \times Y \to \mathbb{R}$ with, for all $(v, \tau, t) \in \mathcal{X}$ and $w^{\mathrm{pw}} \in H(A, \mathcal{T})$,

$$b(v, \tau, t; w^{\mathrm{pw}}) = (\nabla v, \nabla_{NC} w^{\mathrm{pw}})_{L^2(\Omega)} - \omega^2 (\tau, w^{\mathrm{pw}})_{L^2(\Omega)} - \langle t, w^{\mathrm{pw}} \rangle_{\partial\mathcal{T}}.$$

To include the identity $u = \sigma$ and the boundary condition $u = 0$ on $\partial\Omega$, define the reduced ansatz space $X := H^1_0(\Omega) \times H^{-1/2}(\partial\mathcal{T}) \backsimeq \{(v, v) \mid v \in H^1_0(\Omega)\} \times H^{-1/2}(\partial\mathcal{T}) \subset \mathcal{X}$. The bilinear form $b|_{X \times Y}$ reads, for all $(v, t) \in X$ and $w^{\mathrm{pw}} \in H(A, \mathcal{T})$,

$$b(v, t; w^{\mathrm{pw}}) = (\nabla v, \nabla_{NC} w^{\mathrm{pw}})_{L^2(\Omega)} - \omega^2 (v, w^{\mathrm{pw}})_{L^2(\Omega)} - \langle t, w^{\mathrm{pw}} \rangle_{\partial\mathcal{T}}.$$

Theorem 5.1.29 shows that the unique solution $u \in H^1_0(\Omega)$ to (5.83) leads to a unique solution $(u, s) \in X$ to the variational problem

$$b(u, s; w^{\mathrm{pw}}) = (f, w^{\mathrm{pw}})_{L^2(\Omega)} \qquad \text{for all } w^{\mathrm{pw}} \in H^1(\mathcal{T}). \tag{5.84}$$

Set $\mathcal{E}(v, t) := \mathcal{E}_1(v, v, t) \in H(\mathrm{div}, \Omega)$ for all $(v, t) \in X$ with the operator $\mathcal{E}_1$ from Theorem 5.1.31. Let $X_h \subset X$ be a discrete subspace and define the space $Z_h := \{(v_h, \mathcal{E}(v_h, t_h)) \mid$

$(v_h, t_h) \in X_h\} \subset Z := H^1_0(\Omega) \times H(\mathrm{div}, \Omega)$. For all $(v, \vartheta) \in Z$ and $(w, t) \in X$ set

$$\|(v,\vartheta)\|_a^2 := \|\nabla v - \vartheta\|_{L^2(\Omega)}^2 + \|\mathrm{div}\,\vartheta + \omega^2 v\|_{L^2(\Omega)}^2, \tag{5.85a}$$

$$\|(v,\vartheta)\|_Z^2 := \omega^4\|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2 + \|\vartheta\|_{H(\mathrm{div},\Omega)}^2, \tag{5.85b}$$

$$\|(w,t)\|_X^2 := \omega^4\|w\|_{L^2(\Omega)}^2 + \|\nabla w\|_{L^2(\Omega)}^2 + \|\mathcal{E}(w,t)\|_{H(\mathrm{div},\Omega)}^2, \tag{5.85c}$$

$$\|t\|_{H^{-1/2}(\partial\Omega)} := \min\{\|q\|_{H(\mathrm{div},\Omega)} \mid q \in H(\mathrm{div},\Omega) \text{ and } \gamma_\nu^{\mathcal{T}} q = t\} = \|\mathcal{E}(0,t)\|_{H(\mathrm{div},\Omega)}. \tag{5.85d}$$

Let $(u, s) \in X$ solve (5.84). Theorem 5.1.35 shows that the function $(u_h, s_h) \in X_h$ solves the idealized DPG method (5.2) if and only if $(u_h, \mathcal{E}(u_h, s_h)) \in Z_h$ minimizes the functional $\|(u, \mathcal{E}(u, s)) - \bullet\|_a$ over Z_h, that is

$$(u_h, \mathcal{E}(u_h, s_h)) = \operatorname*{arg\,min}_{z_h \in Z_h} \|(u, \mathcal{E}(u, s)) - z_h\|_a. \tag{5.86}$$

Lemma 5.2.1 (Upper bound for $\|\mathcal{E}(v,0)\|_{H(\mathrm{div},\Omega)}$)**.** *Define the piecewise constant function $h_{\mathcal{T}} \in \mathbb{P}_0(\mathcal{T})$ with $h_{\mathcal{T}}|_T = \mathrm{diam}(T)$ for all $T \in \mathcal{T}$. The function $\mathcal{E}(v,0) \in H(\mathrm{div},\Omega)$ satisfies*

$$\|\mathcal{E}(v,0)\|_{H(\mathrm{div},\Omega)} \le \pi^{-1}(\omega^2+1)\|h_{\mathcal{T}}\nabla v\|_{L^2(\Omega)} \qquad \text{for all } v \in H^1_0(\Omega).$$

Proof. Let $v \in H^1(\Omega)$ and $T \in \mathcal{T}$. Define the L^2 orthogonal projection $\Pi_0 : L^2(T) \to L^2(T)$ onto the space of constant functions $\mathbb{P}_0(T)$, that is, $\Pi_0 = |T|^{-1}\int_T \bullet\,\mathrm{d}x$. Theorem 5.1.31*(ii)* shows that the trace $\gamma_\nu^T \mathcal{E}(v,0) = 0$ and so an integration by parts reveals $\int_T \mathrm{div}\,\mathcal{E}(v,0)\,\mathrm{d}x = 0$. This implies the orthogonality

$$(\Pi_0 v, \mathrm{div}\,\mathcal{E}(v,0))_{L^2(T)} = 0. \tag{5.87}$$

Since $T \in \mathcal{T}$ is convex, the Poincaré inequality $\|(1-\Pi_0)w\|_{L^2(T)} \le \pi^{-1}\mathrm{diam}(T)\|\nabla w\|_{L^2(T)}$ applies for all $w \in H^1(T)$ [PW60, Eq. 3.11 and 4.3]. This inequality, Theorem 5.1.31*(vii)*, an integration by parts, (5.87), and the Cauchy-Schwarz inequality imply

$$\begin{aligned}\|\mathcal{E}(v,0)\|_{H(\mathrm{div},T)}^2 &= (\nabla v, \mathcal{E}(v,0))_{L^2(T)} - \omega^2(v, \mathrm{div}\,\mathcal{E}(v,0))_{L^2(T)}\\ &= -(\omega^2+1)((1-\Pi_0)v, \mathrm{div}\,\mathcal{E}(v,0))_{L^2(T)}\\ &\le \pi^{-1}(\omega^2+1)\,\mathrm{diam}(T)\,\|\nabla v\|_{L^2(T)}\|\mathrm{div}\,\mathcal{E}(v,0)\|_{L^2(T)}.\end{aligned}$$

□

Remark 5.2.2 (Equivalent product norm)**.** *Most DPG methods utilize product norms like*

$$\omega^4\|v\|^2 + \|\nabla v\|^2 + \|t\|_{H^{-1/2}(\partial\Omega)}^2 \qquad \text{for all } (v,t) \in X.$$

Theorem 5.1.31 and Corollary 5.1.32 show, for all $(v,t) \in X$,

$$\begin{aligned}\|\mathcal{E}(v,t)\|_{H(\mathrm{div},\Omega)}^2 &= \|\mathcal{E}(v,0)\|_{H(\mathrm{div},\Omega)}^2 + \|t\|_{H^{-1/2}(\partial\mathcal{T})}^2,\\ \|\mathcal{E}(v,0)\|_{H(\mathrm{div},\Omega)}^2 &\le \omega^4\|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2.\end{aligned}$$

This implies the equivalence of the product norm and the norm $\|\bullet\|_X$ in X. More precisely, there exists a constant $c(\mathcal{T}) \in [1/2, 1)$ such that, for all $(v,t) \in X$,

$$c(\mathcal{T})\,\|(v,t)\|_X^2 \le \omega^4\|v\|^2 + \|\nabla v\|^2 + \|t\|_{H^{-1/2}(\partial\mathcal{T})}^2 \le \|(v,t)\|_X^2. \tag{5.88}$$

Lemma 5.2.1 proves that the equivalence constant $c(\mathcal{T}) \nearrow 1$ as the maximal mesh-size $h_{\max}(\mathcal{T}) := \max\{\mathrm{diam}(T) \mid T \in \mathcal{T}\}$ tends to zero.

The first main result of this section is the computation of the inf-sup and continuity constants β and $\|b\|$ from (5.4). Due to (5.51) this computation reduces to the computation of the ellipticity constants of a LSFEM. Section 3.2.2 provides these constants.

Theorem 5.2.3 (Inf-sup and continuity constant)**.** *The constants β and $\|b\|$ in (5.4) depend in the Dirichlet eigenvalues $\lambda_1 \le \lambda_2 \le \dots$ of the Laplace operator $-\Delta$ from Theorem 2.2.1 and read*

$$\beta^2 := \min_{j\in\mathbb{N}} 1 - \sqrt{\frac{\lambda_j(\omega^2+1)^2}{(\lambda_j+1)(\omega^4+\lambda_j)}} \le \beta(\mathcal{T})^2 := \inf_{(v,t)\in X\setminus\{0\}} \frac{\|b(v,t;\bullet)\|_{Y^*}^2}{\|(v,t)\|_X^2}$$
$$\le \sup_{(v,t)\in X\setminus\{0\}} \frac{\|b(v,t;\bullet)\|_{Y^*}^2}{\|(v,t)\|_X^2} =: \|b(\mathcal{T})\|^2 \le \max_{j\in\mathbb{N}} 1 + \sqrt{\frac{\lambda_j(\omega^2+1)^2}{(\lambda_j+1)(\omega^4+\lambda_j)}} =: \|b\|^2.$$

The estimate is sharp in the sense that for any sequence $(\mathcal{T}_\ell)_{\ell\in\mathbb{N}}$ of partitions with (5.52)–(5.53) and (5.56)–(5.57) the constants $\beta(\mathcal{T}_\ell) \searrow \beta$ and $\|b(\mathcal{T}_\ell)\| \nearrow \|b\|$ as $\ell\to\infty$.

Proof. Theorem 5.1.43 and the computation of the ellipticity constants in Section 3.2.2 (with Theorem 3.1.3, Theorem 3.2.4, and Theorem 3.2.6) imply this theorem. □

Remark 5.2.4 (Inf-sup constant for Poisson)**.** *Theorem 5.2.3 shows that the inf-sup constant $\beta^2 = 1-(1+\lambda_1^{-1})^{-1/2}$ for the Poisson model problem. Let $\Omega=(0,1)^2$ be the unit square domain, then $\lambda_1 = 2\pi^2$. The identities in Example 5.1.13 show that the estimates $\beta_{\rm CDG}$ and $\beta_{\rm CP}$ of the inf-sup constant with the splitting lemma from [CDG16, Thm. 3.3] and [CP18, Thm. 3.3] result in*

$$\beta_{\rm CDG} = 0.441 < \beta_{\rm CP} = 0.607 < \beta = 0.883 \le \inf_{(v,t)\in X\setminus\{0\}} \frac{\|b(v,t;\bullet)\|_{Y(\mathcal{T})^*}}{(\|\nabla v\|_{L^2(\Omega)}^2 + \|t\|_{H^{-1/2}(\partial\mathcal{T})}^2)^{1/2}}.$$

The remainder of this section proves asymptotic exactness properties for a sequence of nested partitions $(\mathcal{T}_\ell)_{\ell\in\mathbb{N}}$ of Ω with the properties from (5.52)–(5.57). To emphasize the dependency on the given partition $\mathcal{T}=\mathcal{T}_\ell$ with $\ell\in\mathbb{N}$, denote the spaces $X_h(\mathcal{T}_\ell) := X_h \subset X(\mathcal{T}_\ell) := X$ and $Y_h(\mathcal{T}_\ell) := Y_h \subset Y(\mathcal{T}_\ell) := Y$, the norms $\|\bullet\|_{X(\mathcal{T}_\ell)} := \|\bullet\|_X$ and $\|\bullet\|_{Y(\mathcal{T}_\ell)} := \|\bullet\|_Y$. Let the bilinear form $b : X\times Y\to\mathbb{R}$ read $b_\ell : X(\mathcal{T}_\ell)\times Y(\mathcal{T}_\ell)\to\mathbb{R}$, and let the operator $\mathcal{E}: X\to H(\mathrm{div},\Omega)$ read $\mathcal{E}_\ell : X(\mathcal{T}_\ell)\to H(\mathrm{div},\Omega)$.

Remark 5.2.5 (Inf-sup constant for the product norm)**.** *If the mesh-size $\max\{\mathrm{diam}(T) \mid T\in\mathcal{T}_\ell\}$ tends to zero as ℓ tends to infinity, the combination of Theorem 5.2.3 and the equivalence of norms in Remark 5.2.2 proves*

$$\inf_{(v,t)\in X(\mathcal{T}_\ell)\setminus\{0\}} \frac{\|b_\ell(v,t;\bullet)\|_{Y(\mathcal{T}_\ell)^*}^2}{\omega^4\|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2 + \|t\|_{H^{-1/2}(\partial\mathcal{T}_\ell)}^2} \searrow \beta \qquad \text{as } \ell\to\infty.$$

Suppose the discrete space $X_h(\mathcal{T}_\ell)$ splits into $X_h(\mathcal{T}_\ell) = V_h(\mathcal{T}_\ell)\times \gamma_\nu^{\mathcal{T}_\ell} W_h(\mathcal{T}_\ell)$ with discrete subspaces $V_h(\mathcal{T}_\ell)\subset H_0^1(\Omega)$ and $W_h(\mathcal{T}_\ell)\subset H(\mathrm{div},\Omega)$ for all $\ell\in\mathbb{N}$. Assume the density property (5.61) with respect to the norm $\|\bullet\|_Z$ from (5.85b), that is, for all $(v,\vartheta)\in Z = H_0^1(\Omega)\times H(\mathrm{div},\Omega)$

$$\lim_{\ell\to\infty} \min_{(v_h,\vartheta_h)\in V_h(\mathcal{T}_\ell)\times W_h(\mathcal{T}_\ell)} \|(v,\vartheta)-(v_h,\vartheta_h)\|_Z = 0. \tag{5.89}$$

Then Theorem 5.1.45 implies the density of $Z_h(\mathcal{T}_\ell) := \{(v_h, \mathcal{E}(v_h, t_h)) \mid (v_h, t_h) \in X_h(\mathcal{T}_\ell)\} \subset Z$ with $\ell \in \mathbb{N}$, that is,

$$\lim_{\ell\to\infty} \min_{(v_h,\mathcal{E}(v_h,t_h))\in Z_h(\mathcal{T}_\ell)} \|(v,\vartheta) - (v_h, \mathcal{E}(v_h, t_h))\|_Z = 0 \qquad \text{for all } (v,\vartheta) \in Z. \tag{5.90}$$

Let $a(\bullet,\bullet)$ and $(\bullet,\bullet)_Z$ induce the norms $\|\bullet\|_a$ and $\|\bullet\|_Z$ from (5.85).

Theorem 5.2.6 (Asymptotic properties of idealized DPG)**.** *Let $\mathfrak{u}_\ell \in X(\mathcal{T}_\ell)$ solve (5.84) and let $\mathfrak{u}_\ell \neq \mathfrak{u}_{h,\ell} \in X_h(\mathcal{T}_\ell) := V_h(\mathcal{T}_\ell) \times \gamma_\nu^{\mathcal{T}_\ell} W_h(\mathcal{T}_\ell)$ solve the idealized DPG method (5.2) for all $\ell \in \mathbb{N}$. Suppose the density property (5.89).*
(i) It holds

$$\lim_{\ell\to\infty} \frac{\|b_\ell(\mathfrak{u}_{h,\ell}, \bullet) - (f,\bullet)_{L^2(\Omega)}\|_{Y(\mathcal{T}_\ell)^*}}{\|\mathfrak{u}_\ell - \mathfrak{u}_{h,\ell}\|_{X(\mathcal{T}_\ell)}} = 1. \tag{5.91}$$

(ii) Let $\mathfrak{u}_{\mathrm{best},\ell} = \arg\min_{x_h\in X_h(\mathcal{T}_\ell} \|\mathfrak{u}_\ell - x_h\|_{X(\mathcal{T}_\ell)}$ be the best approximation of $\mathfrak{u}_\ell \in X(\mathcal{T}_\ell)$ in $X_h(\mathcal{T}_\ell)$ for all $\ell \in \mathbb{N}$, then

$$\lim_{\ell\to\infty} \frac{\|\mathfrak{u}_{\mathrm{best},\ell} - \mathfrak{u}_{h,\ell}\|_{X(\mathcal{T}_\ell)}}{\|\mathfrak{u}_\ell - \mathfrak{u}_{\mathrm{best},\ell}\|_{X(\mathcal{T}_\ell)}} = 0. \tag{5.92}$$

Proof. Theorem 3.2.6 validates the hypotheses (**H1**)–(**H4**) from Section 3.1.1 for $a(\bullet,\bullet)$ and $(\bullet,\bullet)_Z$. Thus, the characterization of the solution in (5.86), the density property (5.90), Theorem 3.1.7–3.1.8, and Remark 3.1.9 result in the theorem. □

Remark 5.2.7 (Asymptotic properties of practical DPG)**.** *If the distance $\|\mathfrak{u}_h^i - \mathfrak{u}_h^p\|_X$ of the solution $\mathfrak{u}_h^i \in X_h$ to the idealized (5.2) and the solution $\mathfrak{u}_h^p \in X_h$ to the practical DPG method is of higher order in the sense of (5.13), the best approximation property (5.92) extends to the practical DPG method (5.3). If in addition the upper bound (5.10b) is of higher-order, the asymptotic exactness (5.91) of the residual $\|b(\mathfrak{u}_h^p, \bullet) - (f,\bullet)_{L^2(\Omega)}\|_{Y_h^*}$ extends to the practical DPG method.*

Numerical experiments

This section concludes with a numerical investigation of the DPG method for the Helmholtz equation. The experiments utilize the finite element spaces from (3.40). More precisely, given a regular triangulation $\mathcal{T}$ of the domain $\Omega \subset \mathbb{R}^2$, the discrete spaces read

$$X_h = X_h(k) = S_0^k(\mathcal{T}) \times \gamma_\nu^{\mathcal{T}} RT_{k-1}(\mathcal{T}) \text{ and } Y_h = Y_h(k+\delta) = \mathbb{P}_{k+\delta}(\mathcal{T}) \text{ with } k \in \mathbb{N}, \delta \in \mathbb{N}_0.$$

Remark 5.2.8 (Existence of discrete solutions)**.** *It is known [GQ14] that the practical DPG method with discrete test space $Y_h = Y_h(k+\delta)$, $k \in \mathbb{N}$ and $\delta \geq 2$, allows for the design of an operator $P : Y \to Y_h$ with (5.5). This implies well-posedness of the practical DPG method. The well-posedness of the practical DPG method is unclear for discrete test spaces $Y_h = Y_h(k+\delta)$ with $k \in \mathbb{N}$ and $\delta < 2$. Computations suggest that the practical DPG method is well-posed for $k = 1$ and $\delta = 0, 1$ as well as $k = 2$ and $\delta = 1$, but not for $k = 2$ and $\delta = 0$.*

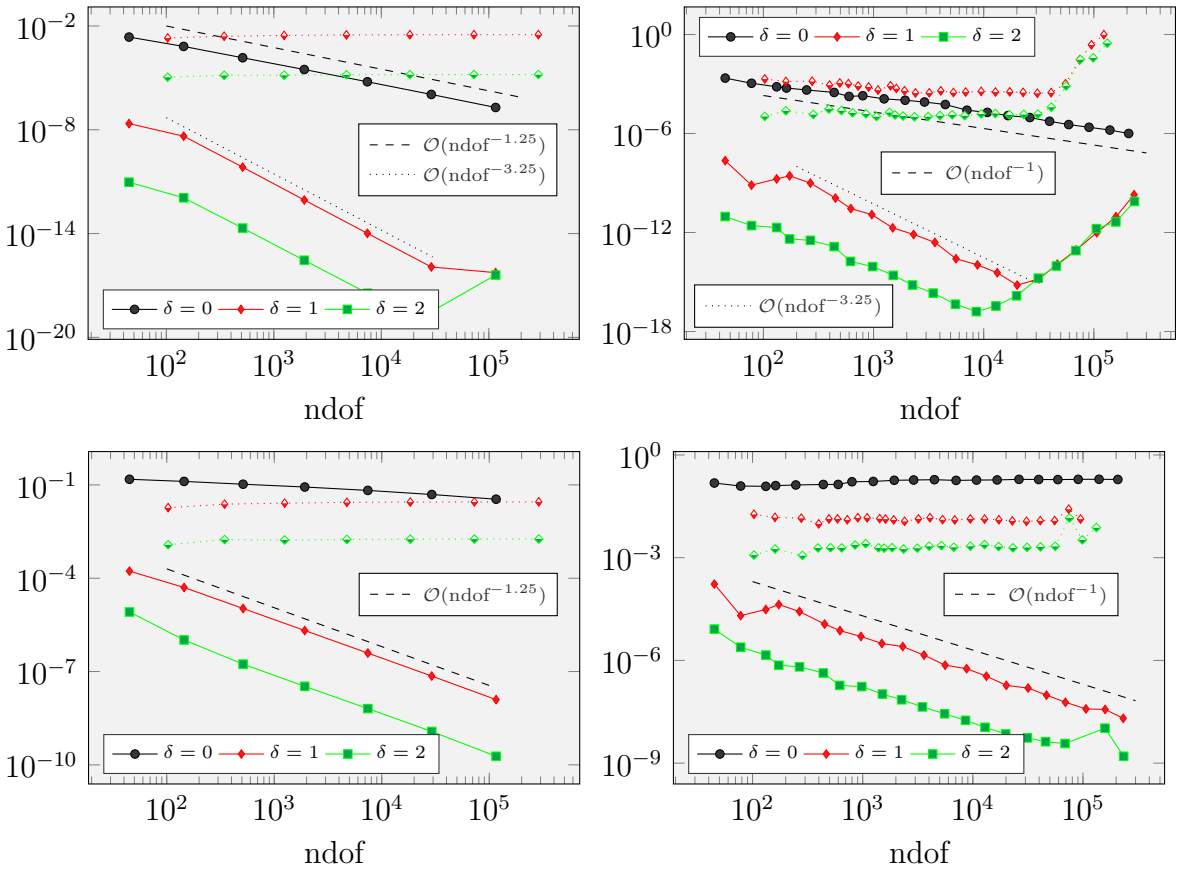


Figure 5.1: Relative distances (5.94a) (top) and (5.94b) (bottom) with polynomial degrees $k = 1$ (solid line) and $k = 2$ (dotted line) and uniform mesh refinements (left) and adaptive mesh refinements (right) in Experiment 1 with frequency $\omega = 1$

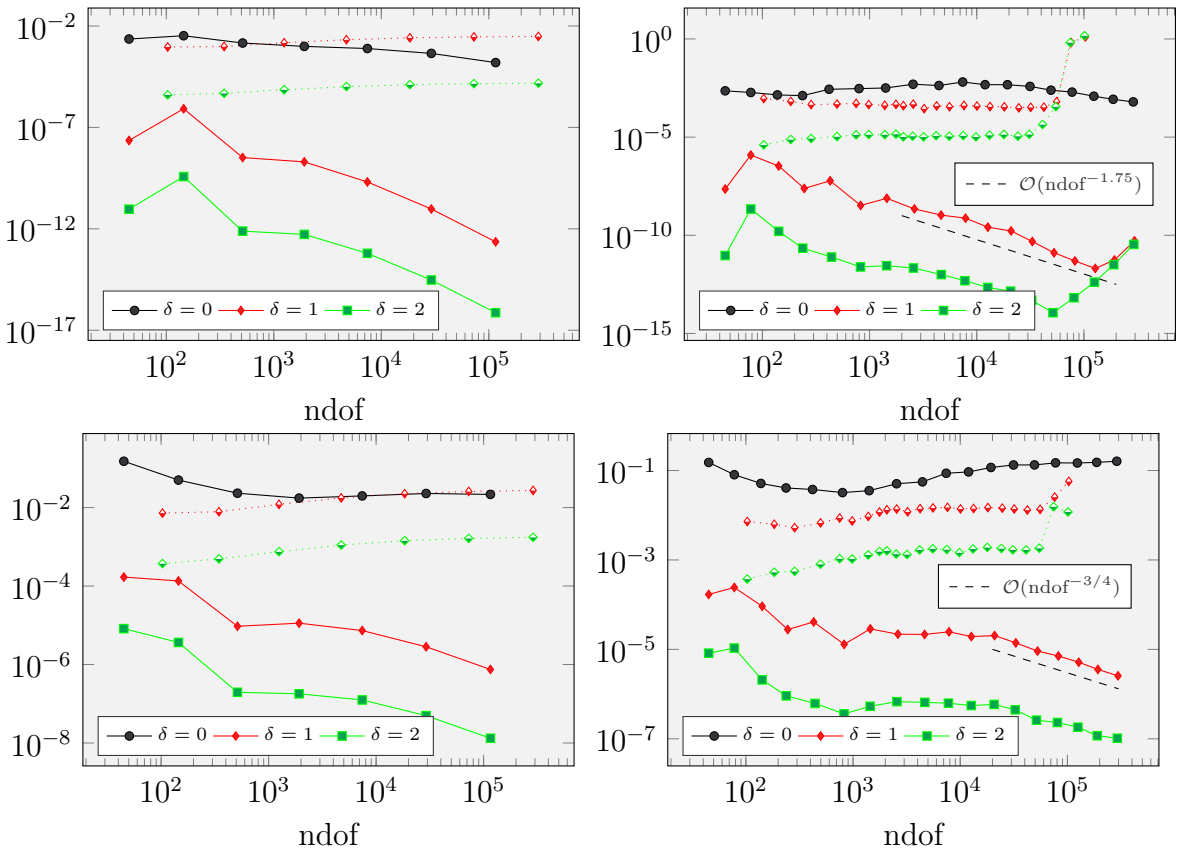


Figure 5.2: Relative distances (5.94a) (top) and (5.94b) (bottom) with polynomial degrees $k = 1$ (solid line) and $k = 2$ (dotted line) and uniform mesh refinements (left) and adaptive mesh refinements (right) in Experiment 1 with frequency $\omega = 3$

Experiment 1 (Practical vs. idealized DPG). This thesis focuses on the analysis of the idealized DPG method (5.2). Since this method is in general not implementable, the first experiment explores the distance of the solution $\mathfrak{u}_h^i \in X_h$ to the idealized (5.2) and the solution $\mathfrak{u}_h^p \in X_h$ to the practical DPG method. More precisely, it investigates the distance $\|\mathfrak{u}_h^i - \mathfrak{u}_h^p\|_X$ and the difference $\|b(\mathfrak{u}_h^p, \bullet) - (f, \bullet)_{L^2(\Omega)}\|_{Y^*}^2 - \|b(\mathfrak{u}_h^p, \bullet) - (f, \bullet)_{L^2(\Omega)}\|_{Y_h^*}^2$ for the Helmholtz equation.

Given a regular triangulation $\mathcal{T}$ of the L-shaped domain $\Omega = (-1,1)^2 \setminus [0,1)^2$, the experiment computes the solution $\mathfrak{u}_h(k,\delta) \in X_h(k) = S_0^k(\mathcal{T}) \times RT_{k-1}(\mathcal{T}) \subset X$ to the practical DPG method with broken discrete test space $Y_h(k+\delta) = \mathbb{P}_{k+\delta}(\mathcal{T}) \subset Y$ for $\delta \in \mathbb{N}_0$, polynomial degrees $k = 1, 2$, right-hand side $F = (f, \bullet)_{L^2(\Omega)} \in Y^*$ with $f \equiv 1$, and frequencies $\omega = 1$ and $\omega = 3$. In other words,

$$\mathfrak{u}_h(k,\delta) = \operatorname*{arg\,min}_{x_h \in X_h(k)} \|b(x_h, \bullet) - F\|_{Y_h(k+\delta)^*} \qquad \text{for all } \delta \in \mathbb{N}_0 \text{ and } k = 1, 2.$$

Theorem 5.1.9 proves that the increasing sequence $(\|b(\mathfrak{u}_h(k,\delta), \bullet) - F\|_{Y_h(k+\delta)^*})_{\delta \in \mathbb{N}}$ converges towards the residual $\|b(\mathfrak{u}_h(k,\infty), \bullet) - F\|_{Y^*} = \|b(\mathfrak{u}_h(k,\infty) - \mathfrak{u}, \bullet)\|_{Y^*}$ with the solution $\mathfrak{u}_h(k,\infty) \in X_h(k)$ to the idealized DPG method (5.2) and the exact solution $\mathfrak{u} \in X$ to (5.84) for $k = 1, 2$. Moreover, Theorem 5.1.9 proves that the error $\|\mathfrak{u}_h(k,\infty) - \mathfrak{u}_h(k,\delta)\|_X \to 0$ as $\delta \to \infty$ and, for all $\delta \in \mathbb{N}_0$ and $k = 1, 2$,

$$\begin{aligned} \|b(\mathfrak{u}_h(k,\infty) - \mathfrak{u}_h(k,\delta), \bullet)\|_{Y^*}^2 &= \|b(\mathfrak{u}_h(k,\delta), \bullet) - F\|_{Y^*}^2 - \|b(\mathfrak{u}_h(k,\infty), \bullet) - F\|_{Y^*}^2 && \text{(5.93a)} \\ &\le \|b(\mathfrak{u}_h(k,\delta), \bullet) - F\|_{Y^*}^2 - \|b(\mathfrak{u}_h(k,\delta), \bullet) - F\|_{Y_h^*}^2. && \text{(5.93b)} \end{aligned}$$

The experiment approximates the norm $\|\bullet\|_{Y^*}$ by the seminorm $\|\bullet\|_{Y_h(6)^*}$ and the solution to the idealized DPG method $\mathfrak{u}_h(k,\infty)$ by $\mathfrak{u}_h(k, 6-k)$ with $k = 1, 2$. Figure 5.1–5.2 plot the relative distances

$$\frac{\|b(\mathfrak{u}_h(k,\delta), \bullet) - F\|_{Y_h(6)^*}^2 - \|b(\mathfrak{u}_h(k,\delta), \bullet) - F\|_{Y_h(k+\delta)^*}^2}{\|b(\mathfrak{u}_h(k,\delta) - \mathfrak{u}, \bullet)\|_{Y_h(6)^*}^2} \qquad \text{and} \qquad \text{(5.94a)}$$

$$\frac{\|b(\mathfrak{u}_h(k, 6-k) - \mathfrak{u}_h(k,\delta), \bullet)\|_{Y_h(6)^*}^2}{\|b(\mathfrak{u}_h(k,\delta) - \mathfrak{u}, \bullet)\|_{Y_h(6)^*}^2} \qquad \text{for } \delta = 0, 1, 2 \text{ and } k = 1, 2. \qquad \text{(5.94b)}$$

The number ndof $:= \dim X_h(k)$. The underlying triangulations $\mathcal{T}$ of the domain Ω into triangles result from uniform (left-hand side) and adaptive (right-hand side) mesh refinements. The adaptively refined meshes result from Algorithm 3 on page 137 with bulk parameter $\Theta = 0.3$ and refinement indicator $(\eta^2(T))_{T \in \mathcal{T}}$ with $\eta^2(T) := \|\eta_h(k,2)\|_{L^2(T)}^2 + \|\nabla \eta_h(k,2)\|_{L^2(T)}^2$ for all $T \in \mathcal{T}$ and the Riesz representation $\eta_h(k,2) \in Y_h(k+2)$ with

$$(\eta_h(k,2), y_h)_Y = b(\mathfrak{u}_h(k,2), y_h) - F(y_h) \qquad \text{for all } y_h \in Y_h(k+2) \text{ and } k = 1, 2.$$

The numerator in (5.94b) approximates the error (5.93a), the numerator in (5.94a) approximates the upper bound (5.93b), and the denominator $\|b(\mathfrak{u}_h(k,\delta) - \mathfrak{u}, \bullet)\|_{Y_h(6)^*}^2$ approximates the error $\|b(\mathfrak{u}_h(k,\delta) - \mathfrak{u}, \bullet)\|_{Y^*}^2$. The experiment leads the following three observations.

1. The distance $\|b(\mathfrak{u}_h(k,\infty) - \mathfrak{u}_h(k,\delta), \bullet)\|_{Y^*}^2$ is of magnitudes smaller than the error $\|b(\mathfrak{u}_h(k,\delta) - \mathfrak{u}, \bullet)\|_{Y^*}^2$ for all $k = 1, 2$ and $\delta = 0, 1, 2$, even in pre-asymptotic regimes

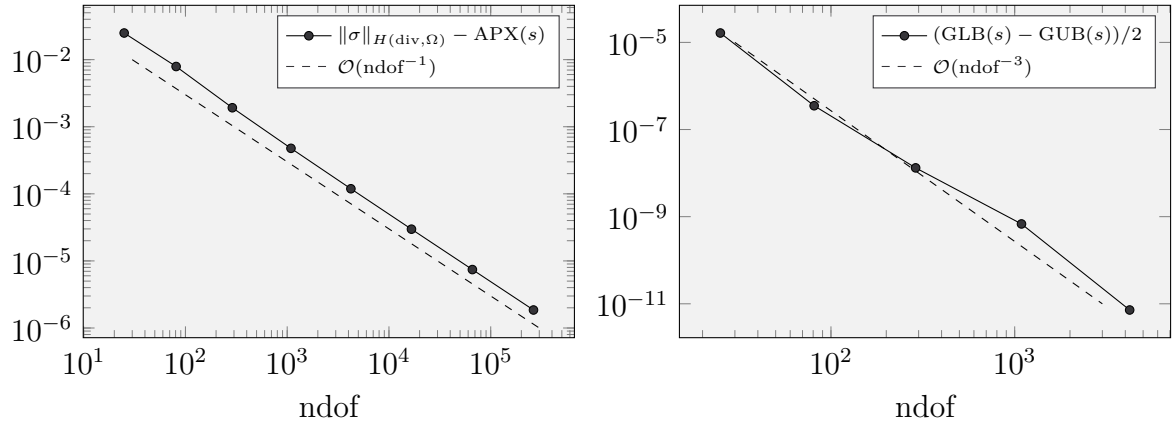


Figure 5.3: Distance between $\|\sigma\|_{H(\mathrm{div},\Omega)}$ and the (approximated) trace norm $\mathrm{APX}(s)$ with $s = \gamma_\nu^{\mathcal{T}}\sigma$ (left) and the (maximal) approximation error $\|s\|_{H^{-1/2}(\partial\mathcal{T})} - \mathrm{APX}(s) \leq (\mathrm{GUB}(s) - \mathrm{GLB}(s))/2$ (right) in Experiment 2

where the DPG method struggles (Experiment 4 shows a pre-asymptotic regime for the DPG method with $\omega = 3$). This suggests that the practical DPG method leads almost to the same result as the idealized DPG method and so justifies the investigation of the idealized DPG method.

2. The error $\|b(\mathfrak{u}_h(k,\infty) - \mathfrak{u}_h(k,\delta), \bullet)\|_{Y^*}^2$ is significantly smaller than the upper bound $\|b(\mathfrak{u}_h(k,\delta), \bullet) - F\|_{Y^*}^2 - \|b(\mathfrak{u}_h(k,\delta), \bullet) - F\|_{Y_h^*}^2$ from (5.93b). Moreover, the rate of convergence for the relative error (5.94b) can be larger than the rate of convergence of (5.94a). This indicates that the upper bound in Theorem 5.1.9 is too pessimistic.

3. The relative error (5.94b) tends to zero for $k = 1$ and $\delta = 0, 1, 2$. This suggest that the asymptotic exactness result (5.92) extends to the practical DPG method with $k = 1$ and $\delta = 0, 1, 2$. The asymptotic exactness result (5.91) extends to the practical DPG method with $k = 1$ and $\delta = 1, 2$, but not to $k = 1$ and $\delta = 0$ since the relative error (5.94a) does not tend to zero. The relative error (5.94a) remains almost constant for $k = 2$. This suggests that the asymptotic exactness results do not apply to the practical DPG method with higher-order ansatz space $X_h(k)$, $k > 1$.

Experiment 2 (Trace norm). This experiment investigates the trace norm $\|s\|_{H^{-1/2}(\partial\mathcal{T})}$ for a sequence of uniformly refined regular triangulations $\mathcal{T}$ of the unit square domain $\Omega = (0,1)^2$ and the trace

$$s = \gamma_\nu^{\mathcal{T}}\sigma \qquad \text{with} \qquad \sigma(x,y) := \nabla x(1-x)y(1-y) \text{ for all } (x,y) \in \Omega.$$

Corollary 5.1.32 motivates the following two-sided bounds.

1. Minimize the norm of all functions in a subset $W_h \subset H(\mathrm{div},\Omega)$ with trace s, that is,

$$\begin{aligned}\|s\|_{H^{-1/2}(\partial\mathcal{T})} &= \min\{\|\vartheta\|_{H(\mathrm{div},\Omega)} \mid \vartheta \in H(\mathrm{div},\Omega) \text{ with } \gamma_\nu^{\mathcal{T}}\vartheta = s\} \\ &\leq \min\{\|\vartheta_h\|_{H(\mathrm{div},\Omega)} \mid \vartheta_h \in W_h \text{ with } \gamma_\nu^{\mathcal{T}}\vartheta_h = s\} =: \mathrm{GUB}(s).\end{aligned}$$

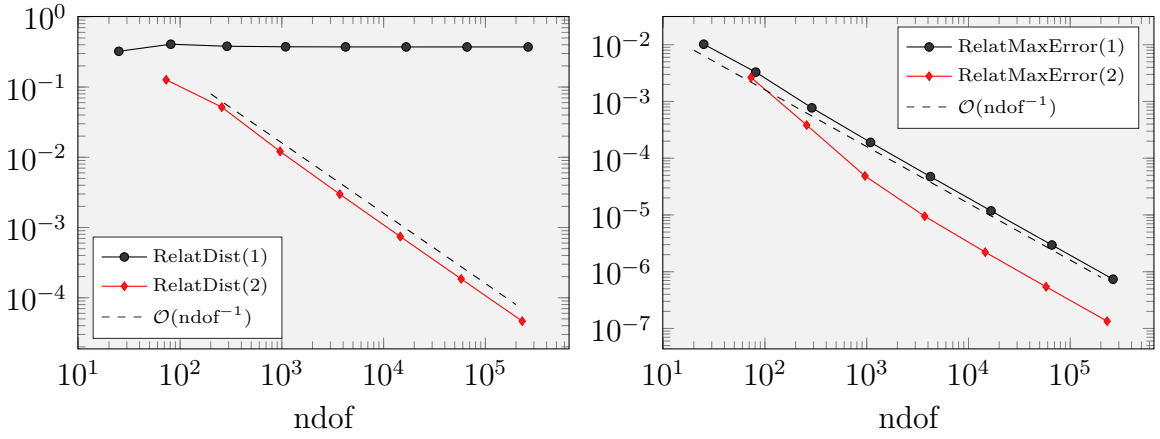


Figure 5.4: Relative distance RelatDist from (5.95a) of the $H(\mathrm{div})$ norm and approximated trace norm (left) and the relative maximal error RelatMaxError from (5.95b) of the guaranteed bounds (right) in Experiment 2

2. Define a discrete subspace $Y_h \subset Y$ and compute the dual norm

$$\mathrm{GLB}(s) := \|b(0,s;\bullet)\|_{Y_h^*} \leq \|b(0,s;\bullet)\|_{Y^*} = \|s\|_{H^{-1/2}(\partial\mathcal{T})}.$$

The computation of the guaranteed lower bound $\mathrm{GLB}(s)$ in this experiment utilizes the space $Y_h := \mathbb{P}_3(\mathcal{T})$. The discrete solution $\sigma_h(T) \in RT_2(T) \cap H_0(\mathrm{div},T)$ to the variational problem $(\sigma_h(T),\vartheta_h)_{H(\mathrm{div},T)} = (\sigma,\vartheta_h)_{H(\mathrm{div},T)}$ for all $\vartheta_h \in RT_2(T) \cap H_0(\mathrm{div},T)$ and $T \in \mathcal{T}$ results in the upper bound

$$\mathrm{GUB}(s)^2 = \sum_{T\in\mathcal{T}} \|\sigma - \sigma_h(T)\|^2_{H(\mathrm{div},T)} \leq \|\sigma\|^2_{H(\mathrm{div},\Omega)}.$$

The left-hand side of Figure 5.3 displays the distance of $\|\sigma\|_{H(\mathrm{div},\Omega)}$ and the mean value

$$\mathrm{APX}(s) := (\mathrm{GUB}(s) + \mathrm{GLB}(s))/2 \leq \|\sigma\|_{H(\mathrm{div},\Omega)}.$$

The right-hand side displays the maximal approximation error $\|s\|_{H^{-1/2}(\partial\mathcal{T})} - \mathrm{APX}(s) \leq (\mathrm{GUB}(s) - \mathrm{GLB}(s))/2$. The approximation error is of magnitudes smaller than the distance $\|\sigma\|_{H(\mathrm{div},\Omega)} - \mathrm{APX}(s)$ (for $\mathrm{ndof} := \dim X_h \geq 16641$ the approximation error is tiny and so numerical difficulties result in $\mathrm{GUB} \leq \mathrm{GLB}$). The approximation error satisfies

$$\|s\|_{H^{-1/2}(\partial\mathcal{T})} - \mathrm{GLB}(s) = \mathcal{O}(\mathrm{ndof}^{-3}) \quad \text{and} \quad \mathrm{GUB}(s) - \|s\|_{H^{-1/2}(\partial\mathcal{T})} = \mathcal{O}(\mathrm{ndof}^{-3}).$$

The small approximation error and $\|\sigma\|_{H(\mathrm{div},\Omega)} - \mathrm{APX}(s) = \mathcal{O}(\mathrm{ndof}^{-1})$ indicate

$$\|\sigma\|_{H(\mathrm{div},\Omega)} - \|s\|_{H^{-1/2}(\partial\mathcal{T})} = \mathcal{O}(\mathrm{ndof}^{-1}).$$

This moderate speed of convergence raises the question if the approximation of the trace norm by the $H(\mathrm{div},\Omega)$ norm (as for example in [CGHW14, CH16, CP18]) allows for sufficiently accurate approximations of the error $\|\mathfrak{u} - \mathfrak{u}_h\|_X$ with the solution $\mathfrak{u}_h \in X_h \subset X$ to the DPG method. Figure 5.4 answers this question with an investigation of trace

norms of the interpolation error $e_k := \sigma - \mathcal{I}_k\sigma$ with the interpolation operator $\mathcal{I}_k : H(\mathrm{div},\Omega) \cap H^{1/2+\varepsilon}(\Omega;\mathbb{R}^d) \to RT_{k-1}(\mathcal{T})$ for $\varepsilon > 0$ and polynomial degree $k = 1, 2$ from [Mon03, Thm. 5.25]. More precisely, it displays the relative distances

$$\mathrm{RelatDist}(k) := (\|e_k\|_{H(\mathrm{div},\Omega)} - \mathrm{APX}(\gamma_\nu^{\mathcal{T}} e_k))/\|e_k\|_{H(\mathrm{div},\Omega)}, \tag{5.95a}$$

$$\mathrm{RelatMaxError}(k) := (\mathrm{GUB}(\gamma_\nu^{\mathcal{T}} e_k) - \mathrm{GLB}(\gamma_\nu^{\mathcal{T}} e_k))/\|e_k\|_{H(\mathrm{div},\Omega)}. \tag{5.95b}$$

The computation of the guaranteed lower bound $\mathrm{GLB}(\gamma_\nu^{\mathcal{T}} e_k)$ with $k = 1, 2$ utilizes the space $Y_h := \mathbb{P}_{k+2}(\mathcal{T})$ and the guaranteed upper bound $\mathrm{GUB}(\gamma_\nu^{\mathcal{T}} e_k)^2 = \sum_{T\in\mathcal{T}} \|e_k - e_{h,k}(T)\|^2_{H(\mathrm{div},\Omega)}$ with $e_{h,k}(T) \in RT_{k+1}(T) \cap H_0(\mathrm{div}, T)$ and

$$(e_{h,k}(T), \vartheta_h)_{H(\mathrm{div},T)} = (e_k, \vartheta_h)_{H(\mathrm{div},T)} \qquad \text{for all } \vartheta_h \in RT_{k+1}(T) \cap H_0(\mathrm{div}, T).$$

The maximal approximation error $(\mathrm{GUB}(\gamma_\nu^{\mathcal{T}} e_k) - \mathrm{GLB}(\gamma_\nu^{\mathcal{T}} e_k))/2$ with $k = 1, 2$ is of magnitudes smaller than the distance $\|e_k\|_{H(\mathrm{div},\Omega)} - \mathrm{APX}(\gamma_\nu^{\mathcal{T}} e_k)$ and so the left-hand side of Figure 5.4 indicates

$$\begin{aligned}
&(\|e_1\|_{H(\mathrm{div},\Omega)} - \|\gamma_\nu^{\mathcal{T}} e_1\|_{H^{-1/2}(\partial\mathcal{T})})/\|e_1\|_{H(\mathrm{div},\Omega)} \to 0.3723 \qquad \text{as ndof} \to \infty,\\
&(\|e_2\|_{H(\mathrm{div},\Omega)} - \|\gamma_\nu^{\mathcal{T}} e_2\|_{H^{-1/2}(\partial\mathcal{T})})/\|e_2\|_{H(\mathrm{div},\Omega)} = \mathcal{O}(\mathrm{ndof}^{-1}).
\end{aligned}$$

This observation suggests that the approximation of the trace norm by the $H(\mathrm{div})$ norm in lowest-order methods results in an additional error which is of the same order as the interpolation error (and so the error $\|\mathfrak{u} - \mathfrak{u}_h\|_X$ of the DPG method). Thus, this approximation is sufficient to show rates of convergence but does not allow for asymptotically exact evaluations of the efficiency indices like

$$I_{\mathrm{eff}} := \|b(\mathfrak{u}_h, \bullet) - (f, \bullet)_{L^2(\Omega)}\|_{Y_h^*}/\|\mathfrak{u} - \mathfrak{u}_h\|_X.$$

The distance of the error $\|\gamma_\nu^{\mathcal{T}} e_2\|_{H^{-1/2}(\partial\mathcal{T})}$ in the trace norm and the error $\|e_2\|_{H(\mathrm{div},\Omega)}$ in the $H(\mathrm{div})$ norm is of higher order. However, the computation of the $H(\mathrm{div})$ norm for higher-order DPG methods ($k \geq 2$) requires a non-trivial extension of the trace approximation $s_h \in \gamma_\nu^{\mathcal{T}} RT_{k-1}(\mathcal{T})$ (unlike in the case $k = 1$ due to interior degrees of freedom). In other words, the $H(\mathrm{div})$ norm leads in lowest-order DPG methods immediately to vague approximation of the trace norm, the approximation of the trace norms by the $H(\mathrm{div})$ norm in higher-order DPG methods requires a post-processing, but is more precise.

This thesis recommends the approximation of the error $\|s - s_h\|_{H^{-1/2}(\partial\mathcal{T})}$ in the trace norm with $s \in H^{-1/2}(\partial\mathcal{T})$ and $s_h \in \gamma_\nu^{\mathcal{T}} RT_{k-1}(\mathcal{T})$ by $\mathrm{GLB}(s - s_h) = \|b(0, s - s_h; \bullet)\|_{Y_h^*}$ with $Y_h = \mathbb{P}_{k+\delta}(\mathcal{T})$ and $k, \delta \in \mathbb{N}$. Figure 5.4 shows that this approximation leads to a much smaller error than the approximation of the trace norm by the $H(\mathrm{div})$ norm. Moreover, the computation of $\mathrm{GLB}(s - s_h)$ can reuse the inverse Gram matrix G^{-1} from the computation of the solution $\mathfrak{u}_h \in X_h$ to the DPG method (see Listing A.13 on page 134) and so reduces to a matrix-vector multiplication.

Experiment 3 (Asymptotic exactness). This experiment applies the primal DPG method of this section with discrete spaces $X_h := S^1_0(\mathcal{T}) \times \gamma_\nu^{\mathcal{T}} RT_0(\mathcal{T})$ and $Y_h(1 + \delta) = \mathbb{P}_{1+\delta}(\mathcal{T})$, $\delta \in \mathbb{N}_0$, and the LSFEM from Section 3.2.2 with discrete space $Z_h := S^1_0(\mathcal{T}) \times$

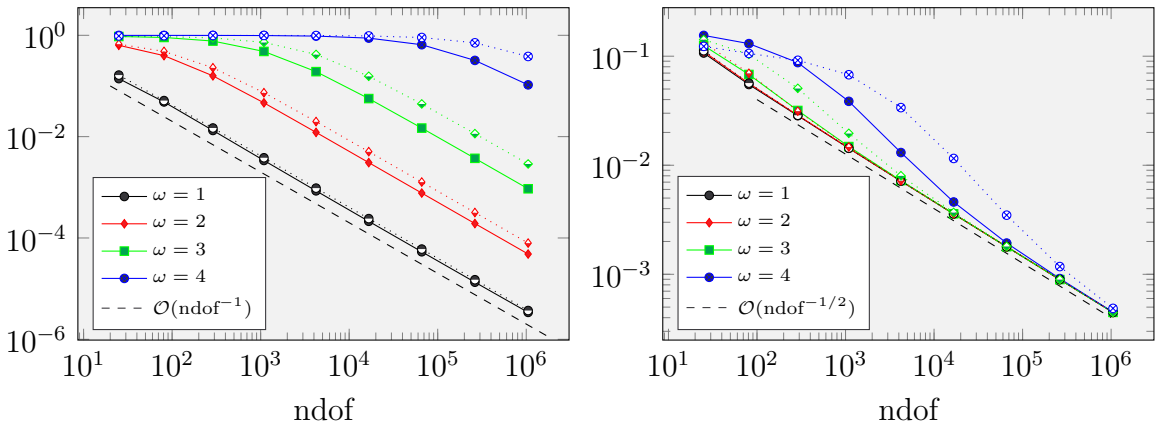


Figure 5.5: The left-hand side plots the distance DistDPG (solid line) and DistLS (dotted line) from (5.99), the right-hand side plots the energy error $\|\nabla(u - u_h^{\mathrm{DPG}}(2))\|_{L^2(\Omega)}$ (solid line) and $\|\nabla(u - u_h^{\mathrm{LS}})\|_{L^2(\Omega)}$ (dotted line) in Experiment 3 with frequencies $\omega = 1, 2, 3, 4$

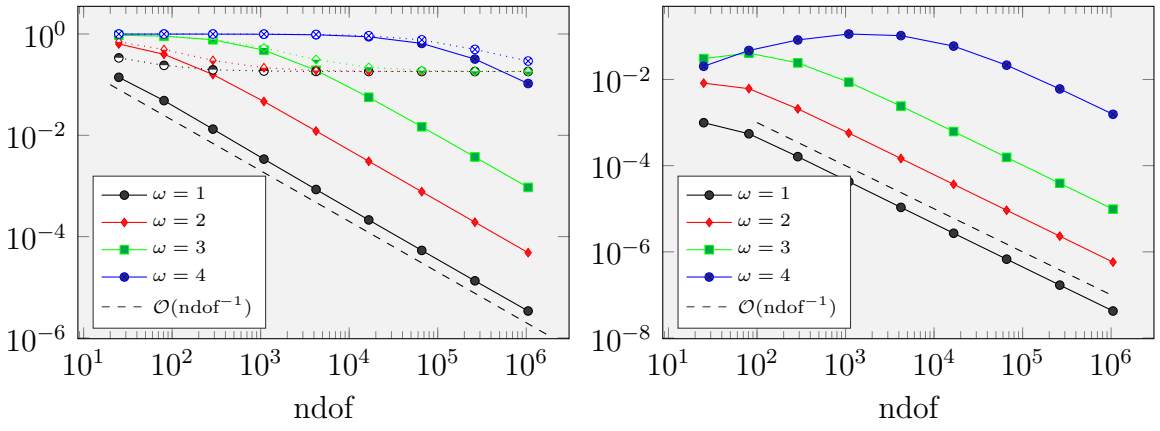


Figure 5.6: The left-hand side plots the distance (5.100) for $\delta = 2$ (solid line) and $\delta = 0$ (dotted line), the right-hand side plots the distance (5.101) in Experiment 3 with frequencies $\omega = 1, 2, 3, 4$

$RT_0(\mathcal{T})$ and least-squares functional $LS(f; v_h, \tau_h) = \|\nabla v_h - \tau_h\|^2_{L^2(\Omega)} + \|\omega^2 v_h + \operatorname{div} \tau_h + f\|^2_{L^2(\Omega)}$ for all $(v_h, \tau_h) \in Z_h$ to the Helmholtz equation (2.7) with known solution

$$u(x,y) = x(1-x)y(1-y) \quad \text{for all } (x,y) \in \Omega = (0,1)^2, \qquad \sigma = \nabla u, \qquad s = \gamma_\nu^{\mathcal{T}} \sigma.$$

In other words, given the right-hand side $f = -\operatorname{div} \sigma - \omega^2 u$, a frequency $\omega > 0$, and a number $\delta \in \mathbb{N}_0$, the experiment computes the solutions to

$$\mathfrak{u}_h^{\mathrm{DPG}}(\delta) = (u_h^{\mathrm{DPG}}(\delta), s_h(\delta)) = \operatorname*{arg\,min}_{x_h \in X_h} \|b(x_h, \bullet) - (f, \bullet)_{L^2(\Omega)}\|_{Y_h(1+\delta)^*}, \tag{5.96a}$$

$$\mathfrak{u}_h^{\mathrm{LS}} = (u_h^{\mathrm{LS}}, \sigma_h) = \operatorname*{arg\,min}_{z_h \in Z_h} LS(f, z_h). \tag{5.96b}$$

Since the discrete space $Z_h \subset Z = H^1_0(\Omega) \times H(\operatorname{div}, \Omega)$ satisfies the density property (5.89) for regular mesh refinements with vanishing maximal mesh-size (cf. [Bra07, Chap. 3.5] and [Bar15, Lem. 3.6]), the asymptotic exactness results from Theorem 5.2.6 and Theorem 3.1.7–3.1.8 apply to the idealized DPG method and the LSFEM. These results imply the convergence of the ratios

$$\frac{LS(f; \mathfrak{u}_h^{\mathrm{LS}})}{\|(u,\sigma) - \mathfrak{u}_h^{\mathrm{LS}}\|^2_Z} \qquad \text{and} \qquad \frac{\|\nabla(u - u_h^{\mathrm{LS}})\|_{L^2(\Omega)}}{\min_{v_h \in S^1_0(\mathcal{T})} \|\nabla(u - v_h)\|_{L^2(\Omega)}} \tag{5.97}$$

and (provided (5.10b) is of higher order)

$$\frac{\|b(\mathfrak{u}_h^{\mathrm{DPG}}(\delta), \bullet) - (f, \bullet)_{L^2(\Omega)}\|^2_{Y_h(1+\delta)^*}}{\|(u,s) - \mathfrak{u}_h^{\mathrm{DPG}}\|^2_X} \qquad \text{and} \qquad \frac{\|\nabla(u - u_h^{\mathrm{DPG}}(\delta))\|_{L^2(\Omega)}}{\min_{v_h \in S^1_0(\mathcal{T})} \|\nabla(u - v_h)\|_{L^2(\Omega)}} \tag{5.98}$$

towards one (with norm $\|\bullet\|_Z$ in the proof of the asymptotic best approximation results, cf. Remark 3.2.8 and the supplementary material of [CS18]). The computation of the norm $\|\bullet\|_X$ utilizes the accurate approximation $\|b(0, s_h(\delta) - \gamma_\nu^{\mathcal{T}} \sigma; \bullet)\|_{Y_h(3)^*}$ with $Y_h(3) = \mathbb{P}_3(\mathcal{T})$ of the trace norm $\|s_h(\delta) - \gamma_\nu^{\mathcal{T}} \sigma\|_{H^{-1/2}(\partial\mathcal{T})}$ (see Experiment 3). Figure 5.5 visualizes the asymptotic exactness result (5.91) for the frequencies $\omega = 1, 2, 3, 4$ and a sequence of uniformly refined triangulations. More precisely, it shows the convergence history plot of

$$\mathrm{DistDPG} := 1 - \frac{\|b(\mathfrak{u}_h^{\mathrm{DPG}}(2), \bullet) - (f, \bullet)_{L^2(\Omega)}\|^2_{Y_h(3)^*}}{\|(u,s) - \mathfrak{u}_h^{\mathrm{DPG}}(2)\|^2_X}, \tag{5.99a}$$

$$\mathrm{DistLS} := 1 - \frac{LS(f; \mathfrak{u}_h^{\mathrm{LS}})}{\|(u,\sigma) - \mathfrak{u}_h^{\mathrm{LS}}\|^2_Z}. \tag{5.99b}$$

The experiment indicates, with $\mathrm{ndof} := \dim X_h$,

$$1 - \frac{LS(f; \mathfrak{u}_h^{\mathrm{LS}})}{\|(u,\sigma) - \mathfrak{u}_h^{\mathrm{LS}}\|^2_Z} = \mathcal{O}(\mathrm{ndof}^{-1}) = 1 - \frac{\|b(\mathfrak{u}_h^{\mathrm{DPG}}(2), \bullet) - (f, \bullet)_{L^2(\Omega)}\|^2_{Y_h(3)^*}}{\|(u,s) - \mathfrak{u}_h^{\mathrm{DPG}}\|^2_X}.$$

As the squared frequency ω^2 approach the first Dirichlet eigenvalue $\lambda_1 = 2\pi^2 \approx 19.7$, a pre-asymptotic regime without convergence of the ratios and poor approximation properties occurs. The numerical experiments in Section 3.2.2 investigate and discuss this phenomenon. The pre-asymptotic regime in the DPG method seems to be smaller than

the pre-asymptotic regime in the LSFEM. Moreover, the energy error in the DPG method is smaller than the energy error in the LSFEM in the pre-asymptotic regime, that is,

$$\|\nabla(u - u_h^{\mathrm{DPG}}(2))\|_{L^2(\Omega)} / \|\nabla(u - u_h^{\mathrm{LS}})\|_{L^2(\Omega)} \ll 1 \qquad \text{for coarse triangulations } \mathcal{T}.$$

Figure 5.6 investigates the DPG method with $\delta = 0$ and $\delta = 2$. The left-hand side displays the convergence history plot of

$$1 - \frac{\|b(\mathfrak{u}_h^{\mathrm{DPG}}(\delta), \bullet) - (f, \bullet)_{L^2(\Omega)}\|^2_{Y_h(\delta)^*}}{\|(u,s) - \mathfrak{u}_h^{\mathrm{DPG}}(\delta)\|_X^2} \qquad \text{with } \delta = 0, 2. \tag{5.100}$$

This distance tends to zero for $\delta = 2$, but remains almost constant for $\delta = 0$. This observation is similar to the observation in Figure 5.1–5.2: The seminorm $\|\bullet\|_{Y_h(0)^*}$ does not allow for a sufficient accurate approximation of the norm $\|\bullet\|_{Y^*}$ and so results in an underestimation of the ratio. This leads to the existence of a constant $\varepsilon > 0$ such that, for all sufficiently fine meshes $\mathcal{T}$,

$$\frac{\|b(\mathfrak{u}_h^{\mathrm{DPG}}(0), \bullet) - (f, \bullet)_{L^2(\Omega)}\|^2_{Y_h(0)^*}}{\|(u,s) - \mathfrak{u}_h^{\mathrm{DPG}}(0)\|_X^2} + \varepsilon < \frac{\|b(\mathfrak{u}_h^{\mathrm{DPG}}(0), \bullet) - (f, \bullet)_{L^2(\Omega)}\|^2_{Y^*}}{\|(u,s) - \mathfrak{u}_h^{\mathrm{DPG}}(0)\|_X^2}.$$

The right-hand side of Figure 5.6 displays the convergence history plot of

$$1 - \frac{\|\nabla(u - u_h^{\mathrm{DPG}}(0))\|_{L^2(\Omega)}}{\|\nabla(u - u_h^{\mathrm{DPG}}(2))\|_{L^2(\Omega)}}. \tag{5.101}$$

Since the error $\|\nabla(u_h - u_h^{\mathrm{DPG}}(\delta))\|_{L^2(\Omega)}$ with $u_h = \arg\min_{x_h \in S^1_0(\mathcal{T})} \|x_h - u\|_{L^2(\Omega)}$ is of higher order (see Theorem 5.2.6 and Experiment 1), (5.101) tends to zero for $\delta = 0, 2$. Surprisingly, the energy error $\|\nabla(u - u_{h,0})\|_{L^2(\Omega)} \leq \|\nabla(u - u_{h,2})\|_{L^2(\Omega)}$ in all computations, that is, the DPG method with test space $Y_h(0)$ results in slightly better approximations than the DPG method with test space $Y_h(2)$.

Experiment 4 (Instant stability). Let the right-hand side $f \equiv 1$, let the frequency $\omega > 0$, and let $\mathcal{T}$ be a regular triangulation of the L-shaped domain $\Omega = (-1,1)^2 \setminus [0,1)^2$. This experiment compares the solution $(u_h^{\mathrm{DPG}}, s_h) \in S^1_0(\mathcal{T}) \times \gamma_\nu^{\mathcal{T}} RT_0(\mathcal{T})$ to the DPG method (5.96a) with $Y_h = \mathbb{P}_3(\mathcal{T})$, the solution $(u_h^{\mathrm{LS}}, \sigma_h) \in S^1_0(\mathcal{T}) \times RT_0(\mathcal{T})$ to the LSFEM (5.96b), and the solution $u_h^{\mathrm{C}} \in S^1_0(\mathcal{T})$ to the Courant FEM

$$(\nabla u_h^{\mathrm{C}}, \nabla v_h)_{L^2(\Omega)} - \omega^2 (u_h^{\mathrm{C}}, v_h)_{L^2(\Omega)} = (f, v_h)_{L^2(\Omega)} \qquad \text{for all } v_h \in S^1_0(\mathcal{T}).$$

Since the exact solution is unknown, the experiment compares error estimators. For the solution to the DPG method and the LSFEM these estimators read $\eta^2_{\mathrm{DPG}} := \|b(u_h^{\mathrm{DPG}}, s_h; \bullet) - (f, \bullet)_{L^2(\Omega)}\|^2_{Y_h^*}$ and $\eta^2_{\mathrm{LS}} := LS(f; u_h^{\mathrm{LS}}, \sigma_h)$. The error estimation for the Courant FEM utilizes the residual-based error estimator from [BHP17, CKNS08, CN12, FFP14]. This estimator involves the diameter $h_T := \mathrm{diam}(T)$ for all $T \in \mathcal{T}$ and the normal jump $[\nabla u_h^{\mathrm{C}} \cdot \nu]|_E := (\nabla u_h^{\mathrm{C}}|_{T_1}) \cdot \nu_{T_1} + (\nabla u_h^{\mathrm{C}}|_{T_2}) \cdot \nu_{T_2}$ for all edges $E = T_1 \cap T_2$ with $T_1, T_2 \in \mathcal{T}$, where ν_{T_1} and ν_{T_2} denote the normal unit vector pointing outward the triangles T_1 and T_2. The estimator reads $\eta^2_{\mathrm{C}} := \sum_{T \in \mathcal{T}} \eta_{\mathrm{C}}(T)^2$ with

$$\eta_{\mathrm{C}}(T)^2 := h_T^2 \|f + \Delta u_h^{\mathrm{C}} + \omega^2 u_h^{\mathrm{C}}\|^2_{L^2(T)} + h_T \|[\nabla u_h^{\mathrm{C}} \cdot \nu]\|^2_{L^2(\partial T \cap \Omega)} \qquad \text{for all } T \in \mathcal{T}.$$

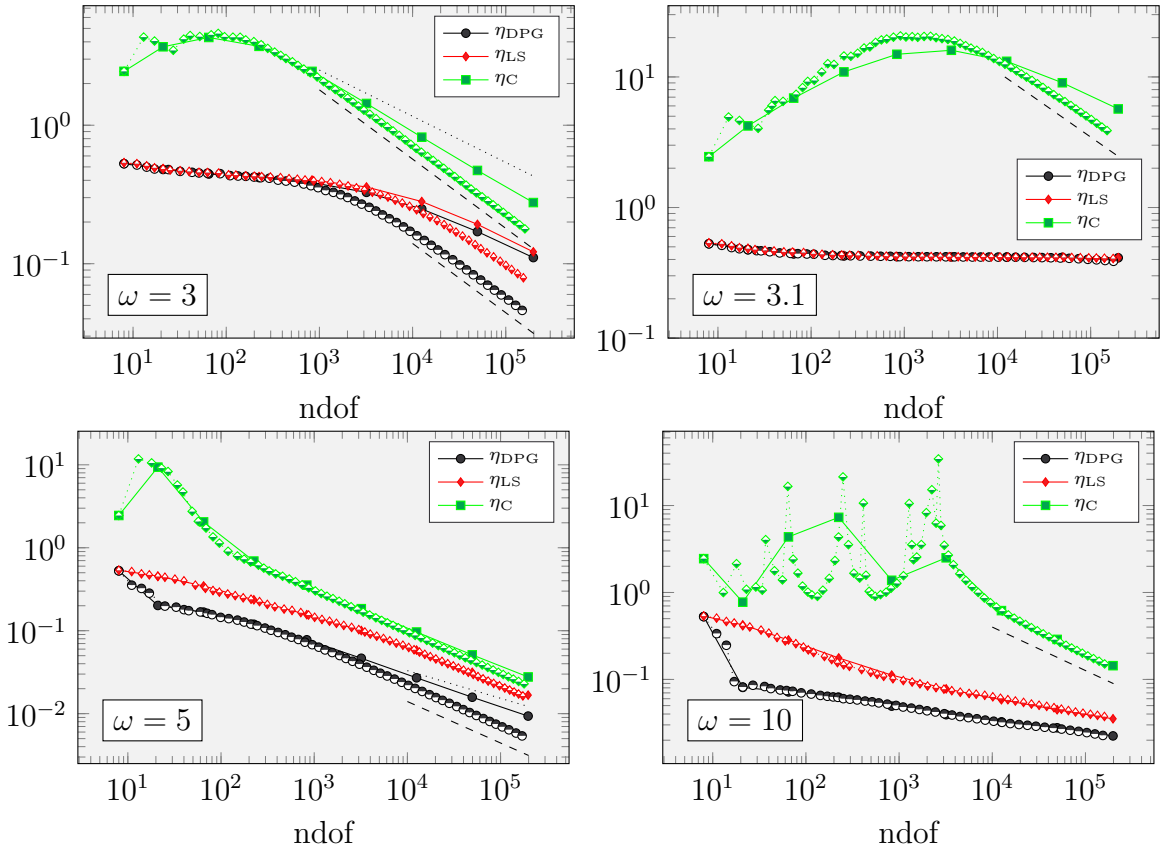


Figure 5.7: Comparison of the error estimators η_{DPG}, η_{LS}, and η_C for the DPG method, LSFEM, and Courant FEM with uniform (filled markers) and adaptive (half-filled markers) mesh refinements for various frequencies ω and the slopes $-1/2$ (- - -) and $-1/3$ (·····) in Experiment 4

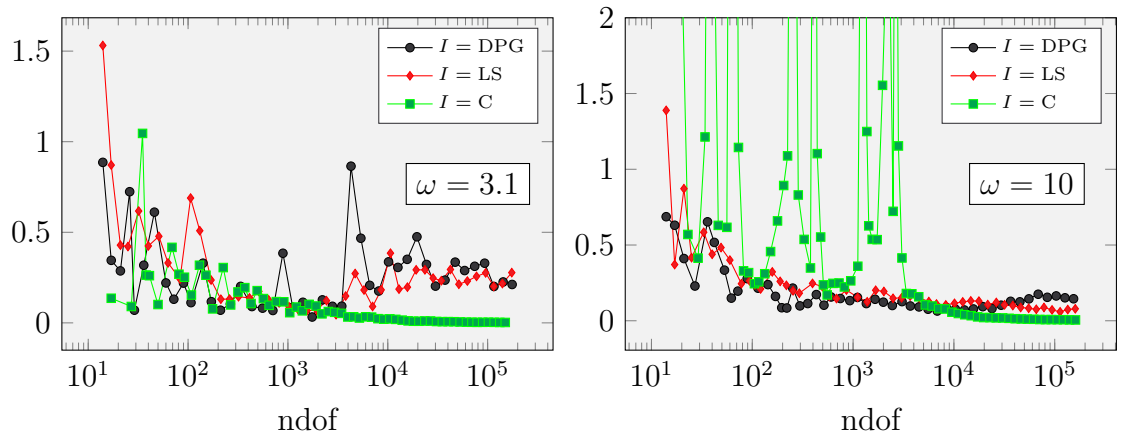


Figure 5.8: Relative changes $\|\nabla(u^I_{\ell+1} - u^I_\ell)\|_{L^2(\Omega)}/\|\nabla u^I_\ell\|_{L^2(\Omega)}$ of the solutions $u^I_\ell \in S^1_0(\mathcal{T}_\ell)$ and $u^I_{\ell+1} \in S^1_0(\mathcal{T}_{\ell+1})$ to the DPG method ($I = \mathrm{DPG}$), LSFEM ($I = \mathrm{LS}$), and Courant FEM ($I = \mathrm{C}$) with respect to the mesh $\mathcal{T}_\ell$ and the adaptively refined mesh $\mathcal{T}_{\ell+1}$

All estimators are equivalent to the error. The mesh $\mathcal{T}$ results from a uniform mesh refinement (filled markers) or an adaptive mesh refinement (half-filled markers). The local contributions of the error estimators η^2_{DPG}, η^2_{LS}, and η^2_{C} drive the adaptive mesh refinement in Algorithm 3 with bulk parameter $\Theta = 0.1$. Figure 5.7 displays the convergence history plot of the error estimators and ndof $= \dim S^1(\mathcal{T})$. It shows pre-asymptotic regimes without good approximations. For large frequencies, this phenomenon is known as pollution [BS97], that is, the solution $u^{\mathrm{C}} \in S^1_0(\mathcal{T})$ to the Courant FEM requires a maximal mesh-size $h_{\max} := \max\{\mathrm{diam}(T) \mid T \in \mathcal{T}\}$ with $h_{\max}\omega^2 < 1$ for good approximations. Moreover, Theorem 5.2.3 and Section 3.2.2 show large inf-sup and continuity constants in the DPG method and LSFEM for large frequencies and frequencies close to an eigenfrequency. This leads to large constants in the a priori estimates and so indicates poor approximations. Similar arguments lead to large constants in the a priori estimates for the Courant FEM as well (see for example the computation of the (discrete) inf-sup constant in [Bar15, p. 101]). Although all three methods result in poor approximations on coarse meshes, Figure 5.7–5.9 indicate differences:

1. In contrast to the DPG method and the LSFEM, the solutions to the Courant FEM with respect to a triangulation $\mathcal{T}_\ell$ and the refinement $\mathcal{T}_{\ell+1}$ of $\mathcal{T}_\ell$ differ significantly in a pre-asymptotic regime.
2. The error of the Courant FEM seems to be large and does not decrease in a pre-asymptotic regime.
3. The Courant FEM overcomes the pre-asymptotic regime without (optimal) convergence faster than the DPG method and the LSFEM.
4. The adaptive mesh refinement for the LSFEM results in a strong refinement of the boundary $\partial\Omega$, the adaptive mesh refinement for the DPG method refines the re-entrant corner, and the adaptive mesh refinement for the Courant FEM results in an almost uniform refinement in the pre-asymptotic regime without (optimal) convergence.

The observations in 1–2 indicate that the DPG method and the LSFEM are stable, that is, the errors do not increase significantly. This motivates adaptive schemes, driven by the discrete solutions on coarse meshes (which violate the criterion $\omega^2 h_{\max} \ll 1$). Indeed, the numerical experiment in [PD17] shows a DPG method with good approximations of a Gaussian beam for adaptively refined meshes that satisfy $1 < \omega^2 h_{\max}$. The Courant FEM is unstable, that is, the error $\|\nabla(u - u_h^C)\|_{L^2(\Omega)}$ oscillates strongly. However, the pre-asymptotic regime is much smaller than the pre-asymptotic regime of the minimal residual methods. Moreover, [BHP17] proves that the adaptive scheme converges with the optimal rate $\|\nabla(u - u_h^C)\|_{L^2(\Omega)} = \mathcal{O}(\mathrm{ndof}^{-1/2})$ in an asymptotic regime (the size of the pre-asymptotic regime strongly depends on the frequency ω).

Discussion. The overall conclusions from the numerical benchmarks in Experiment 1–4 are in agreement with the theoretical predictions of this work. Experiment 1–2 show very small differences of the idealized and practical DPG method, that is, the error $\|\mathfrak{u}_h^i - \mathfrak{u}_h^p\|_X$ of the solution $\mathfrak{u}_h^i \in X_h$ to the idealized and $\mathfrak{u}_h^p \in X_h$ to the practical DPG method is much smaller than the error $\|\mathfrak{u} - \mathfrak{u}_h^p\|_X$ with the exact solution $\mathfrak{u} \in X$ and the approximation of the norm $\|\bullet\|_{Y^*}$ by the seminorm $\|\bullet\|_{Y_h^*}$ results in good approximations, especially for

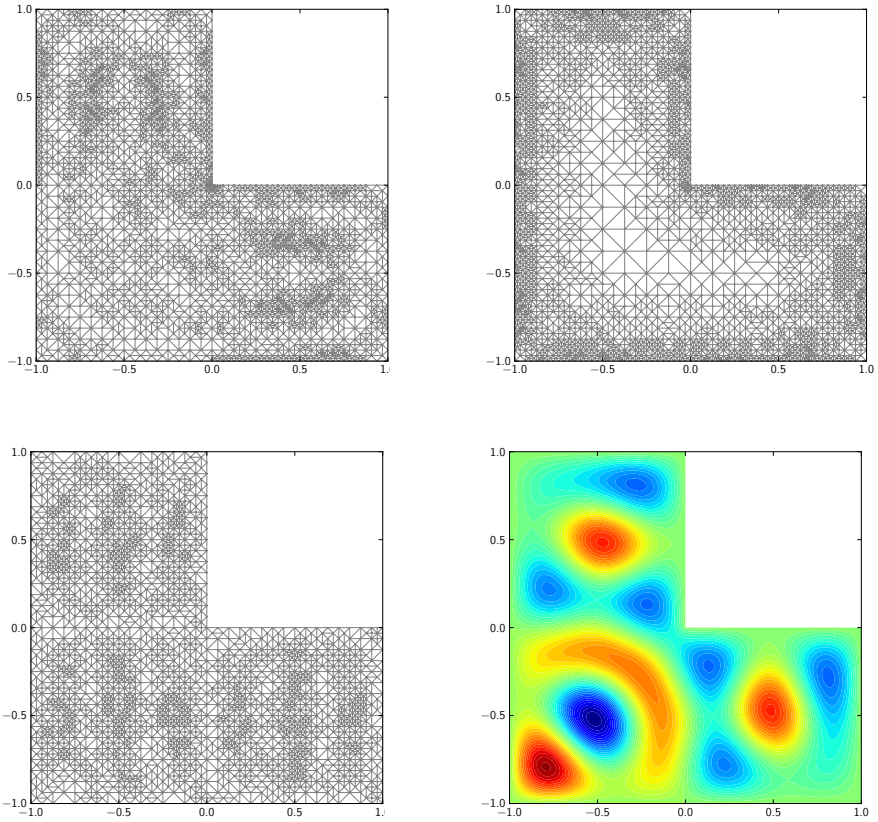


Figure 5.9: Adaptively refined meshes with the DPG method (top left), LSFEM (top right), and Courant FEM (bottom left) after ndof = $\dim S^1(\mathcal{T})$ exceeds 2500 and a contour plot of the (approximated) solution

discrete test spaces Y_h of higher polynomial degrees. This indicates that the practical DPG method with higher-order test spaces behaves very similar to the idealized DPG method and so justifies the investigation of the latter (in general not implementable) approach.

Experiment 2 recommends the approximation of the trace norm $\|s\|_{H^{-1/2}(\partial\mathcal{T})}$ with $s \in H^{-1/2}(\partial\mathcal{T})$ (or other minimal extension norms) by the (cheap) computation of the residual $\|b(0,s;\bullet)\|_{Y_h^*}$ instead of the (more common) evaluation of the $H(\mathrm{div},\Omega)$ norm.

The idealized approach is related to a least-squares method. The relation leads to the question: Why to use the DPG method instead of the LSFEM? Besides several practical aspects (for example easy to implement ultra-weak formulations), Experiment 3–4 indicate a simple answer: The DPG method seems to perform better than the LSFEM, that is, errors in a pre-asymptotic regime are smaller. Surprisingly, the DPG method seems to perform even better with low-order test spaces, that is, the error of the DPG method with low-order test space is smaller than the error of the DPG method with higher-order test space in pre-asymptotic regimes. A justification of the latter two observations requires a more detailed study of the DPG method and the LSFEM for the Helmholtz equation.

A further open topic is the comparison of adaptive minimal residual methods and the Courant FEM for large frequencies. Experiment 4 with $\omega = 10$ indicates that the adaptive Courant FEM performs better. However, uniform mesh refinements seem to result in optimal rates of convergence as well. This indicates that adaptive mesh refinement is unnecessary for this problem (at least in the regime with ndof $< 2 \times 10^5$). A more detailed comparison of the adaptive LSFEM, DPG method, and Courant FEM requires a problem where adaptively refined meshes improve the results with uniformly refined meshes significantly (as for example a Gaussian beam).

5.2.2 Elasticity

Given a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ with $d \in \mathbb{N}$, Lamé parameters λ and μ, the fourth-order elasticity tensor $\mathbb{C} = \mathbb{C}(\lambda,\mu)$ from (2.11), the symmetric gradient $\varepsilon(\bullet) = (\nabla\bullet + (\nabla\bullet)^\top)/2$, and a body force $f \in L^2(\Omega;\mathbb{R}^d)$, linear elasticity seeks the (weak) solution $u \in H^1_0(\Omega;\mathbb{R}^d)$ to

$$-\mathrm{div}\,\mathbb{C}\varepsilon(u) = f. \tag{5.102}$$

Many finite element methods struggle in the incompressible limit $\lambda \to \infty$, see for example [BS92a, BS92b]. This phenomenon is known as locking. Remedies are mixed methods (see [CDFH00] and the references therein), non-conforming methods (see [CF01b] and the references therein), and least-squares finite element methods (see [CS04]). This section utilizes the framework of Section 5.1.4 to design a primal DPG method which is related to the locking-free LSFEM from [CS04]. This relation proves robustness in the incompressible limit $\lambda \to \infty$.

There are several ways of defining operators A^* and B such that (5.102) equals the abstract problem $ABu = f$ from (5.29). The arguments from Section 5.2.1 prove that the design of Section 5.1.4 with operators

$$A^* := -\mathrm{div}\,\mathbb{C}^{1/2}, \qquad B := \mathbb{C}^{1/2}\varepsilon(\bullet), \qquad A = \mathbb{C}^{1/2}\nabla \qquad \text{and} \qquad C = 0 \tag{5.103}$$

leads to an asymptotically exact DPG method. However, numerical experiments indicate locking, that is, a large Lamé parameter λ results in a large pre-asymptotic regime without

improvements of the discrete solution (see for example Figure 5.10 at the end of this section). Therefore, set the operators

$$A^* := -\operatorname{div}\mathbb{C}, \qquad B := \varepsilon(\bullet), \qquad A := \mathbb{C}\nabla, \qquad \text{and} \qquad C = 0. \tag{5.104}$$

For all Lipschitz domains $\omega \subset \Omega$ the domains of the operators in (5.104) read

$$\begin{aligned}\operatorname{dom}(A) &= H(A,\omega) = \operatorname{dom}(B) = H(B,\omega) = H^1(\omega;\mathbb{R}^d),\\ \operatorname{dom}(A^*) &= H(A^*,\omega) = \{\vartheta \in L^2(\omega;\mathbb{R}^{d\times d}) \mid \operatorname{div}\mathbb{C}\vartheta \in L^2(\omega;\mathbb{R}^d)\} = \mathbb{C}^{-1}H(\operatorname{div},\omega;\mathbb{R}^{d\times d})\\ &=: H(\operatorname{div}\mathbb{C},\omega;\mathbb{R}^{d\times d}).\end{aligned}$$

The properties of the operators $-\operatorname{div}$, ∇, and $\mathbb{C}$ imply Assumption 5.1.21–5.1.22 and so the design from Section 5.1.4 applies. The definition in Lemma 5.1.25 and the component-wise application of the trace operators from Theorem 5.1.15 prove for all Lipschitz domains $\omega \subset \Omega$ and functions $v \in H^1(\omega;\mathbb{R}^d)$, $\vartheta \in H(\operatorname{div}\mathbb{C},\omega;\mathbb{R}^{d\times d})$ the identity

$$\langle \gamma^\omega_{A^*}\vartheta, \gamma^\omega_A v\rangle_{\partial\omega} := (\mathbb{C}\nabla v,\vartheta)_{L^2(\omega)} + (v,\operatorname{div}\mathbb{C}\vartheta)_{L^2(\omega)} = \langle \gamma^\omega_\nu \mathbb{C}\vartheta, \gamma^\omega_0 v\rangle_{H^{-1/2}(\partial\omega),H^{1/2}(\partial\omega)}.$$

Let $\mathcal{T}$ be a partition of the domain Ω into a finite number of disjoint and non-empty Lipschitz domains with (5.25). The trace space reads, for all $T \in \mathcal{T}$,

$$\Gamma_{A^*}(\partial T) = \gamma^T_{A^*} H(\operatorname{div}\mathbb{C},T;\mathbb{R}^{d\times d}) = \gamma^T_\nu H(\operatorname{div},T;\mathbb{R}^{d\times d}) =: H^{-1/2}(\partial T,\mathbb{C};\mathbb{R}^d).$$

The trace operator (5.41) on the skeleton reads

$$\gamma^{\mathcal{T}}_{A^*} : H(\operatorname{div}\mathbb{C},\Omega;\mathbb{R}^{d\times d}) \to \Gamma_{A^*}(\partial\mathcal{T}) =: H^{-1/2}(\partial\mathcal{T},\mathbb{C};\mathbb{R}^d) \subset \prod_{T\in\mathcal{T}} H^{-1/2}(\partial T,\mathbb{C};\mathbb{R}^d)$$

and maps all $\vartheta \in H(\operatorname{div}\mathbb{C},\Omega;\mathbb{R}^{d\times d})$ onto $\gamma^{\mathcal{T}}_{A^*}\vartheta = (\gamma^T_{A^*}\vartheta|_T)_{T\in\mathcal{T}}$. To include the boundary condition, define the reduced ansatz space

$$\begin{aligned}X &:= H^1_0(\Omega;\mathbb{R}^d) \times H^{-1/2}(\partial\mathcal{T};\mathbb{R}^d) \simeq H^1_0(\Omega;\mathbb{R}^d) \times \{0\} \times H^{-1/2}(\partial\mathcal{T};\mathbb{R}^d)\\ &\subset \mathcal{X} := H^1(\Omega;\mathbb{R}^d) \times \{0\} \times H^{-1/2}(\partial\mathcal{T};\mathbb{R}^d).\end{aligned}$$

The broken test space $Y = H(A,\mathcal{T})$ reads

$$Y := H^1(\mathcal{T};\mathbb{R}^d) := \{w^{\mathrm{pw}} \in L^2(\Omega;\mathbb{R}^d) \mid w^{\mathrm{pw}}|_T \in H^1(T;\mathbb{R}^d) \text{ for all } T \in \mathcal{T}\}.$$

The operator A_{NC} from (5.40) equals $A_{NC} = \mathbb{C}\nabla_{NC}$ with the component-wise application of the gradient $(\nabla_{NC}w^{\mathrm{pw}})|_T = \nabla w^{\mathrm{pw}}|_T$ for all $w^{\mathrm{pw}} \in H^1(\mathcal{T};\mathbb{R}^d)$. The norm in Y reads

$$\|\bullet\|_Y := (\|A_{NC}\bullet\|^2_{L^2(\Omega)} + \|\bullet\|^2_{L^2(\Omega)})^{1/2} = (\|\mathbb{C}\nabla_{NC}\bullet\|^2_{L^2(\Omega)} + \|\bullet\|^2_{L^2(\Omega)})^{1/2}.$$

Define the non-conforming symmetric gradient $\varepsilon_{NC}(\bullet) := (\nabla_{NC}\bullet + (\nabla_{NC}\bullet)^\top)/2$. The bilinear forms in (5.42) equal, for all $(v,t) \in X$ with $t = (t_T)_{T\in\mathcal{T}}$ and $w^{\mathrm{pw}} \in H^1(\mathcal{T};\mathbb{R}^d)$,

$$\begin{aligned}\langle t, w^{\mathrm{pw}}\rangle_{\partial\mathcal{T}} &= \sum_{T\in\mathcal{T}} \langle t_T, \gamma^T_A w^{\mathrm{pw}}|_T\rangle_{\partial T},\\ b(v,t;w^{\mathrm{pw}}) &= (\mathbb{C}\varepsilon(v),\nabla_{NC}w^{\mathrm{pw}})_{L^2(\Omega)} - \langle t, w^{\mathrm{pw}}\rangle_{\partial\mathcal{T}} = (\mathbb{C}\varepsilon(v),\varepsilon_{NC}(w^{\mathrm{pw}}))_{L^2(\Omega)} - \langle t, w^{\mathrm{pw}}\rangle_{\partial\mathcal{T}}.\end{aligned}$$

Theorem 5.1.29 proves that the unique solution $u \in H^1_0(\Omega;\mathbb{R}^d)$ to (5.83) leads to a unique solution $(u,s) \in X$ with $s = \gamma^{\mathcal{T}}_{A^*}\varepsilon(u)$ to the variational problem

$$b(u,s;w^{\mathrm{pw}}) = (f,w^{\mathrm{pw}})_{L^2(\Omega)} \qquad \text{for all } w^{\mathrm{pw}} \in H^1(\mathcal{T};\mathbb{R}^d). \tag{5.105}$$

Set $\mathcal{E}(v,t) := \mathcal{E}_1(v,0,t)$ with $\mathcal{E}_1$ from Theorem 5.1.31 for all $(v,t) \in X$ and define the norm

$$\begin{aligned}\|(v,t)\|_X &:= (\|\varepsilon(v)\|^2_{L^2(\Omega)} + \|\mathcal{E}(v,t)\|^2_{H(A^*,\Omega)})^{1/2} \\ &= (\|\varepsilon(v)\|^2_{L^2(\Omega)} + \|\mathcal{E}(v,t)\|^2_{L^2(\Omega)} + \|\operatorname{div}\mathbb{C}\mathcal{E}(v,t)\|^2_{L^2(\Omega)})^{1/2} \qquad \text{for all } (v,t) \in X.\end{aligned} \tag{5.106}$$

The following theorem proves that the inf-sup and continuity constant β and $\|b\|$ from (5.4) are bounded from below and above by λ-independent constants.

Theorem 5.2.9 (Inf-sup and continuity condition)**.** *There exists a λ-independent constant $0 < \beta$ with*

$$0 < \beta \le \frac{\|b(v,t;\bullet)\|_{Y^*}}{\|(v,t)\|_X} \le \sqrt{2} \qquad \textit{for all } (v,t) \in X \setminus \{0\}.$$

The proof of Theorem 5.2.9 requires the following result from [CS04].

Lemma 5.2.10 (Coercivity constant for a locking-free LSFEM)**.** *Let $Z := H^1_0(\Omega;\mathbb{R}^d) \times H(\operatorname{div},\Omega;\mathbb{R}^{d\times d})$ and recall the λ-dependent operator $\mathbb{C}^{-1}$ from Lemma 2.3.1. There exists a λ-independent constant α with*

$$0 < \alpha \le \inf_{(v,\vartheta)\in Z\setminus\{0\}} \frac{\|\varepsilon(v) - \mathbb{C}^{-1}\vartheta\|^2_{L^2(\Omega)} + \|\operatorname{div}\vartheta\|^2_{L^2(\Omega)}}{\|\varepsilon(v)\|^2_{L^2(\Omega)} + \|\mathbb{C}^{-1}\vartheta\|^2_{L^2(\Omega)} + \|\operatorname{div}\vartheta\|^2_{L^2(\Omega)}}.$$

Proof. Set the space $\Sigma := \{\chi \in H(\operatorname{div},\Omega;\mathbb{R}^{d\times d}) \mid \int_\Omega \operatorname{tr}(\chi)\,\mathrm{d}x = 0\}$. Theorem 3.1 in [CS04] proves the existence of a λ-independent constant c with

$$0 < c \le \inf_{0\neq(v,\vartheta)\in H^1_0(\Omega;\mathbb{R}^d)\times\Sigma} \frac{\|\varepsilon(v) - \mathbb{C}^{-1}\vartheta\|^2_{L^2(\Omega)} + \|\operatorname{div}\vartheta\|^2_{L^2(\Omega)}}{\|\varepsilon(v)\|^2_{L^2(\Omega)} + \|\vartheta\|^2_{L^2(\Omega)} + \|\operatorname{div}\vartheta\|^2_{L^2(\Omega)}}. \tag{5.107}$$

An integration by parts reveals, for all $v \in H^1_0(\Omega;\mathbb{R}^d)$,

$$(\varepsilon(v),\mathbb{C}^{-1}I_{d\times d})_{L^2(\Omega)} = (\nabla v,\mathbb{C}^{-1}I_{d\times d})_{L^2(\Omega)} = (v,\operatorname{div}\mathbb{C}^{-1}I_{d\times d})_{L^2(\Omega)} = 0.$$

This and (5.107) imply

$$\begin{aligned}0 < c &\le \inf_{0\neq(v,\vartheta)\in H^1_0(\Omega;\mathbb{R}^d)\times\Sigma} \frac{\|\varepsilon(v) - \mathbb{C}^{-1}\vartheta\|^2_{L^2(\Omega)} + \|\operatorname{div}\vartheta\|^2_{L^2(\Omega)}}{\|\varepsilon(v)\|^2_{L^2(\Omega)} + \|\vartheta\|^2_{L^2(\Omega)} + \|\operatorname{div}\vartheta\|^2_{L^2(\Omega)}} \\ &= 1 - \sup_{0\neq(v,\vartheta)\in H^1_0(\Omega;\mathbb{R}^d)\times\Sigma} \frac{2(\varepsilon(v),\mathbb{C}^{-1}\vartheta)_{L^2(\Omega)}}{\|\varepsilon(v)\|^2_{L^2(\Omega)} + \|\vartheta\|^2_{L^2(\Omega)} + \|\operatorname{div}\vartheta\|^2_{L^2(\Omega)}} \\ &= 1 - \sup_{0\neq(v,\vartheta)\in Z} \frac{2(\varepsilon(v),\mathbb{C}^{-1}\vartheta)_{L^2(\Omega)}}{\|\varepsilon(v)\|^2_{L^2(\Omega)} + \|\vartheta\|^2_{L^2(\Omega)} + \|\operatorname{div}\vartheta\|^2_{L^2(\Omega)}} \\ &= \inf_{0\neq(v,\vartheta)\in Z} \frac{\|\varepsilon(v) - \mathbb{C}^{-1}\vartheta\|^2_{L^2(\Omega)} + \|\operatorname{div}\vartheta\|^2_{L^2(\Omega)}}{\|\varepsilon(v)\|^2_{L^2(\Omega)} + \|\vartheta\|^2_{L^2(\Omega)} + \|\operatorname{div}\vartheta\|^2_{L^2(\Omega)}}.\end{aligned}$$

The combination with $\|\mathbb{C}^{-1}\vartheta\|_{L^2(\Omega)} \le (2\mu)^{-1}\|\vartheta\|_{L^2(\Omega)}$ for all $\vartheta \in H(\operatorname{div},\Omega;\mathbb{R}^{d\times d})$ from Lemma 2.3.1 concludes the proof. □

Proof of Theorem 5.2.9. Let $(v,t) \in X \setminus \{0\}$ and define $\vartheta := \mathbb{C}^{-1}\mathcal{E}(v,t) \in H(\mathrm{div},\Omega;\mathbb{R}^{d\times d})$. The inequality $\|\varepsilon(v) - \mathbb{C}^{-1}\vartheta\|_{L^2(\Omega)}^2 \leq 2(\|\varepsilon(v)\|_{L^2(\Omega)}^2 + \|\mathbb{C}^{-1}\vartheta\|_{L^2(\Omega)}^2)$, the identity from 5.1.31*(vi)*, the definition of the norm $\|\bullet\|_X$, and Lemma 5.2.10 result in

$$\beta^2 = \alpha \leq \frac{\|b(v,t;\bullet)\|_{Y^*}^2}{\|(v,t)\|_X^2} = \frac{\|\varepsilon(v) - \mathbb{C}^{-1}\vartheta\|_{L^2(\Omega)}^2 + \|\mathrm{div}\,\vartheta\|_{L^2(\Omega)}^2}{\|\varepsilon(v)\|_{L^2(\Omega)}^2 + \|\mathbb{C}^{-1}\vartheta\|_{L^2(\Omega)}^2 + \|\mathrm{div}\,\vartheta\|_{L^2(\Omega)}^2} \leq 2. \qquad \square$$

Let $X_h \subset X$ and $Y_h \subset Y$ be discrete subspaces and assume $X_h = V_h \times \gamma_{A^*}^{\mathcal{T}}\mathbb{C}^{-1}W_h$ with discrete subspaces $V_h \subset H_0^1(\Omega;\mathbb{R}^d)$ and $W_h \subset H(\mathrm{div},\Omega;\mathbb{R}^{d\times d})$. Suppose that for all $\lambda > 0$ exists an operator $P : Y \to Y_h$ and a λ-independent constant $\|P_{\max}\| < \infty$ such that

$$P \text{ satisfies (5.5)} \quad \text{and} \quad \|P\| \leq \|P_{\max}\|. \tag{5.108}$$

Remark 5.2.11 (Annulation operator P for a low-order method)**.** *Let $V_h = S_0^1(\mathcal{T};\mathbb{R}^d)$, $W_h = RT_0(\mathcal{T};\mathbb{R}^{d\times d})$, and $Y_h = \mathbb{P}_1(\mathcal{T};\mathbb{R}^d)$ with the spaces from (3.85). Simple modifications of the arguments from [CBHW18, Prop. 4.5] prove that the component-wise application of the non-conforming interpolation operator $\mathcal{I}_{NC}^{\mathrm{loc}}$ from [CBHW18, Eq. 4.7] results in an λ-independent operator $P := \mathcal{I}_{NC}^{\mathrm{loc}}$ with (5.108) and λ-independent upper bound $\|P_{\max}\|$.*

The following theorem proves that the DPG method of this section is locking-free. More precisely, let $\mathfrak{u} = (u,s) \in X$ and $\mathfrak{u}_h = (u_h,s_h) \in X_h$ be the solution to (5.105) and (5.3), then the energy error $\|\varepsilon(u-u_h)\|_{L^2(\Omega)}$ is bounded by a λ-independent constant times the best approximation error in a λ-independent norm.

Theorem 5.2.12 (A priori estimate)**.** *Suppose (5.108). Recall the norm $\|\bullet\|_X$ from (5.106) and the λ-independent constants $\|P_{\max}\| < \infty$ from (5.108) and β from Lemma 5.2.10. The error of the exact solution $(u,s) \in X$ to (5.105) with stress tensor $\sigma := \mathbb{C}\varepsilon(u) \in H(\mathrm{div},\Omega;\mathbb{R}^{d\times d})$ and the solution $(u_h,s_h) \in X_h$ to the DPG method (5.3) satisfy*

$$\|(u-u_h,s-s_h)\|_X^2 \leq 2\beta^{-2}\|P_{\max}\|^2 \min_{(v_h,\vartheta_h)\in V_h\times W_h} 2\,\|\varepsilon(u-v_h)\|_{L^2(\Omega)}^2 + \|\mathbb{C}^{-1}(\sigma-\vartheta_h)\|_{L^2(\Omega)}^2 + \|\mathrm{div}\,(\sigma-\vartheta_h)\|_{L^2(\Omega)}^2. \tag{5.109}$$

Proof. Let $(v_h,t_h) \in X_h$ and $\vartheta_h \in W_h$ with $\gamma_{A^*}^{\mathcal{T}}\mathbb{C}^{-1}\vartheta_h = t_h$. The identity $s = \gamma_{A^*}^{\mathcal{T}}\varepsilon(u) = \gamma_{A^*}^{\mathcal{T}}\mathbb{C}^{-1}\sigma$ (Theorem 5.1.29) and the properties of the operator $\mathcal{E}$ (Theorem 5.1.31) show

$$\begin{aligned}
\|(u,s)-(v_h,t_h)\|_X^2 &= \|\varepsilon(u-v_h)\|_{L^2(\Omega)}^2 + \|\mathcal{E}(u-v_h,0)\|_{H(A^*,\Omega)}^2 + \|\mathcal{E}(0,s-t_h)\|_{H(A^*,\Omega)}^2 \\
&\leq 2\|\varepsilon(u-v_h)\|_{L^2(\Omega)}^2 + \min\{\|\chi\|_{H(A^*,\Omega)}^2 \mid \chi \in H(\mathrm{div}\,\mathbb{C},\Omega;\mathbb{R}^{d\times d}) \text{ with } \gamma_{A^*}^{\mathcal{T}}\chi = s-t_h\} \\
&\leq 2\|\varepsilon(u-v_h)\|_{L^2(\Omega)}^2 + \|\mathbb{C}^{-1}(\sigma-\vartheta_h)\|_{L^2(\Omega)}^2 + \|\mathrm{div}\,(\sigma-\vartheta_h)\|_{L^2(\Omega)}^2.
\end{aligned} \tag{5.110}$$

Since Theorem 5.1.1 implies $\|(u-u_h,s-s_h)\|_X^2 \leq 2\beta^{-2}\|P_{\max}\|^2 \min_{x_h\in X_h}\|(u,s)-x_h\|_X^2$, Theorem 5.2.9 and (5.110) conclude the proof. $\square$

Remark 5.2.13 (λ-independence in (5.109))**.** *Lemma 2.3.1 proves the λ-independent bound $\|\mathbb{C}^{-1}\bullet\|_{L^2(\Omega)}^2 \leq (2\mu)^{-2}\|\bullet\|_{L^2(\Omega)}^2$ for the λ-dependent term on the right-hand side of (5.109).*

Remark 5.2.14 (Numerical difficulties)**.** *The computation of the solution $(u_h,s_h) \in X_h$ to the DPG method (5.3) requires the inversion of the Gram matrix $G = (G_{jk})_{j,k=1,\dots,N}$ with and $G_{jk} = (y_j,y_k)_Y$ for all $j,k = 1,\dots,N$ and basis $y_1,\dots,y_N$ of Y_h. Since Y_h is broken, a reasonable choice of basis functions reduces this problem to the inversion of small submatrices (see Appendix A.2). Unfortunately, a large Lamé parameter λ results in huge condition numbers for these submatrices and so causes numerical difficulties.*

Numerical experiments

The remainder of this section investigates the locking phenomenon with a numerical experiment from [CH16, Sec. 5.4]. The experiment involves the L-shaped domain $\Omega = (-1,1)^2 \setminus [0,1)^2$ and the right-hand side

$$f(x,y) = \begin{cases} (1,0)^\top & \text{for } 0 \leq xy \text{ and } \max\{|x|,|y|\} \leq 0.5, \\ 0 & \text{else.} \end{cases}$$

Given a regular triangulation $\mathcal{T}$ of Ω, the experiment utilize the low-order spaces $S^1_0(\mathcal{T};\mathbb{R}^2)$ and $RT_0(\mathcal{T};\mathbb{R}^{2\times 2})$ from (3.85) and solves the following problems.

1. Seek the solution $\mathfrak{u}_h^{\mathrm{DPG},1} = (u_h^{\mathrm{DPG},1}, s_h^{\mathrm{DPG},1}) \in X_h := S^1_0(\mathcal{T};\mathbb{R}^2) \times \gamma^{\mathcal{T}}_{A^*}\mathbb{C}^{-1}RT_0(\mathcal{T};\mathbb{R}^{2\times 2})$ to the DPG method of this section with $Y_h := \mathbb{P}_3(\mathcal{T};\mathbb{R}^2)$, that is,

$$\mathfrak{u}_h^{\mathrm{DPG},1} = \operatorname*{arg\,min}_{x_h \in X_h} \|b(x_h,\bullet) - (f,\bullet)_{L^2(\Omega)}\|_{Y_h^*} \text{ with } \|\bullet\|^2_{Y_h} = \|\bullet\|^2_{L^2(\Omega)} + \|\mathbb{C}\nabla_{NC}\bullet\|^2_{L^2(\Omega)}.$$

2. Seek the solution $\mathfrak{u}_h^{\mathrm{DPG},2} = (u_h^{\mathrm{DPG},2}, s_h^{\mathrm{DPG},2}) \in X_h$ to the DPG method with the operators from (5.103) and $Y_h := \mathbb{P}_3(\mathcal{T};\mathbb{R}^2)$, that is,

$$\mathfrak{u}_h^{\mathrm{DPG},2} = \operatorname*{arg\,min}_{x_h \in X_h} \|b(x_h,\bullet) - (f,\bullet)_{L^2(\Omega)}\|_{Y_h^*} \text{ with } \|\bullet\|^2_{Y_h} = \|\bullet\|^2_{L^2(\Omega)} + \|\mathbb{C}^{1/2}\nabla_{NC}\bullet\|^2_{L^2(\Omega)}.$$

3. Seek the solution $\mathfrak{u}_h^{\mathrm{LS}} = (u_h^{\mathrm{LS}}, \sigma_h^{\mathrm{LS}}) \in Z_h := S^1_0(\mathcal{T};\mathbb{R}^2) \times RT_0(\mathcal{T};\mathbb{R}^{2\times 2})$ to the locking-free LSFEM from [CS04], that is,

$$\mathfrak{u}_h^{\mathrm{LS}} = \operatorname*{arg\,min}_{(v_h,\vartheta_h)\in Z_h} \|\varepsilon(v_h) - \mathbb{C}^{-1}\vartheta_h\|^2_{L^2(\Omega)} + \|f + \operatorname{div}\vartheta_h\|^2_{L^2(\Omega)}.$$

4. Seek the solution $u_h^{\mathrm{C}} \in S^1_0(\mathcal{T};\mathbb{R}^2)$ to the Courant-FEM, that is,

$$(\mathbb{C}\varepsilon(u_h^{\mathrm{C}}), \varepsilon(v_h))_{L^2(\Omega)} = (f, v_h)_{L^2(\Omega)} \qquad \text{for all } v_h \in S^1_0(\mathcal{T};\mathbb{R}^2).$$

The triangulations $\mathcal{T}$ result from adaptive mesh refinements with Algorithm 3 on page 137 with bulk parameter $\Theta = 0.3$ and refinement indicator

$$\eta^2(T) := \|\varepsilon(u_h^{\mathrm{LS}}) - \mathbb{C}^{-1}\sigma_h^{\mathrm{LS}}\|^2_{L^2(T)} + \|f + \operatorname{div}\sigma_h^{\mathrm{LS}}\|^2_{L^2(T)} \qquad \text{for all } T \in \mathcal{T}.$$

Figure 5.10 compares the solutions to a reference solution $\mathfrak{u}_{\mathrm{ref}} = (u_{\mathrm{ref}}, \sigma_{\mathrm{ref}}) \in Z_{\mathrm{ref}} := S^3_0(\mathcal{T};\mathbb{R}^2) \times RT_2(\mathcal{T};\mathbb{R}^{2\times 2})$ on the finest triangulation $\mathcal{T}$ with

$$(u_{\mathrm{ref}}, \sigma_{\mathrm{ref}}) = \operatorname*{arg\,min}_{(v_h,\vartheta_h)\in Z_{\mathrm{ref}}} \|\varepsilon(v_h) - \mathbb{C}^{-1}\vartheta_h\|^2_{L^2(\Omega)} + \|f + \operatorname{div}\vartheta_h\|^2_{L^2(\Omega)}.$$

More precisely, the left-hand side of Figure 5.10 displays the convergence history plot of the approximated error $\|\varepsilon(u_{\mathrm{ref}} - u_h)\|_{L^2(\Omega)}$ with $u_h \in \{u_h^{\mathrm{DPG},1}, u_h^{\mathrm{DPG},2}, u_h^{\mathrm{LS}}, u_h^{\mathrm{C}}\}$. The right-hand side of Figure 5.10 displays the convergence history plot of the residuals

$$\begin{aligned}
\eta^{\mathrm{DPG},1} &:= \|b(\mathfrak{u}_h^{\mathrm{DPG},1},\bullet) - (f,\bullet)_{L^2(\Omega)}\|_{Y_h^*} \quad \text{with} \quad \|\bullet\|^2_Y = \|\bullet\|^2_{L^2(\Omega)} + \|\mathbb{C}\nabla_{NC}\bullet\|^2_{L^2(\Omega)}, \\
\eta^{\mathrm{DPG},2} &:= \|b(\mathfrak{u}_h^{\mathrm{DPG},2},\bullet) - (f,\bullet)_{L^2(\Omega)}\|_{Y_h^*} \quad \text{with} \quad \|\bullet\|^2_Y = \|\bullet\|^2_{L^2(\Omega)} + \|\mathbb{C}^{1/2}\nabla_{NC}\bullet\|^2_{L^2(\Omega)}, \\
\eta^{\mathrm{LS}} &:= (\|\varepsilon(u_h^{\mathrm{LS}}) - \mathbb{C}^{-1}\sigma_h^{\mathrm{LS}}\|^2_{L^2(\Omega)} + \|f + \operatorname{div}\sigma_h^{\mathrm{LS}}\|^2_{L^2(\Omega)})^{1/2}.
\end{aligned}$$

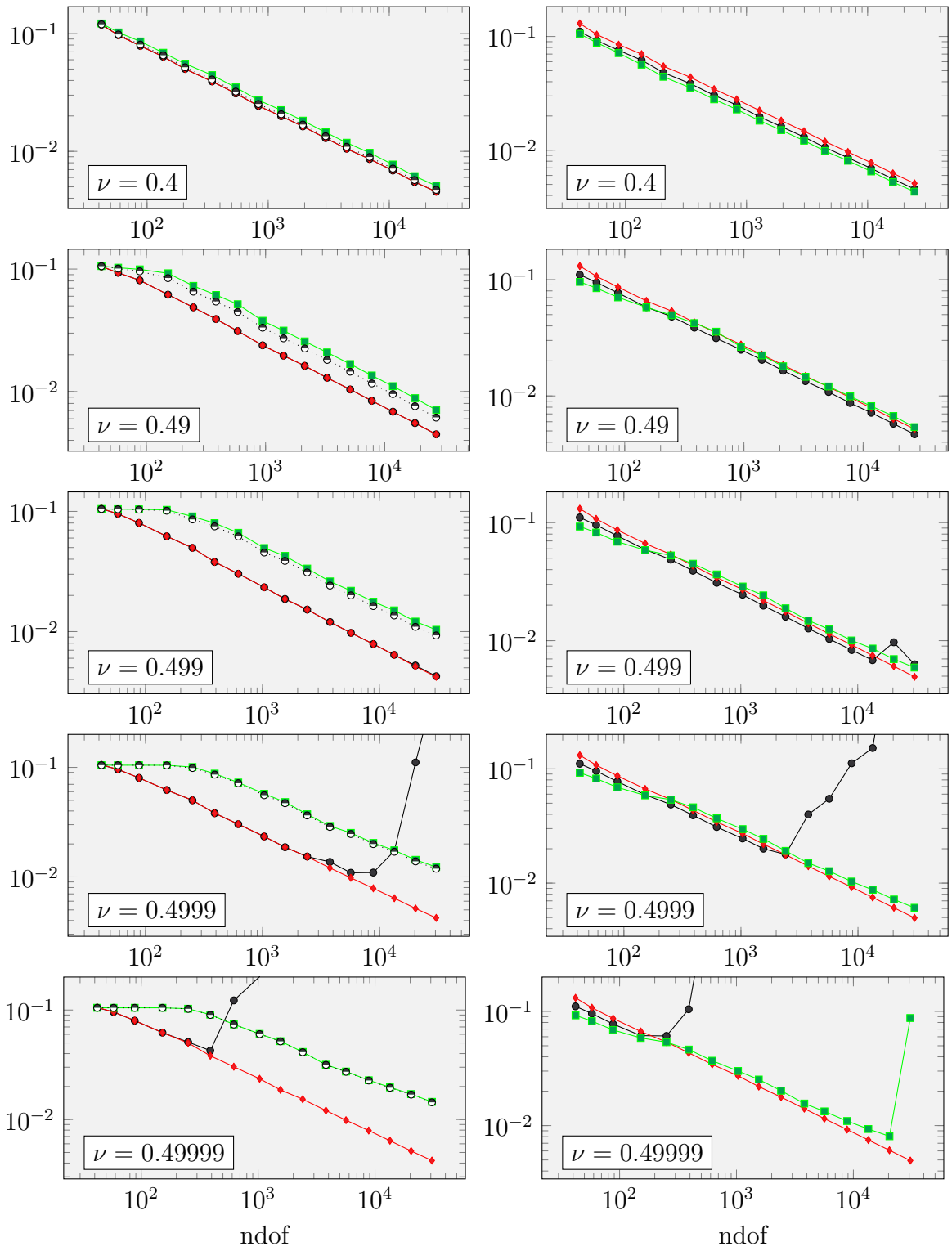


Figure 5.10: The left-hand side displays the error $\|\varepsilon(u_{\text{ref}} - u_h)\|_{L^2(\Omega)}$ with solutions $u_h = u_h^{\text{DPG},1}$ (—●—), $u_h = u_h^{\text{DPG},2}$ (—■—), $u_h = u_h^{\text{LS}}$ (—♦—), and $u_h = u_h^{\text{C}}$ (···○···); the right side displays the residuals η^{LS} (—♦—), $\eta^{\text{DPG},1}$ (—●—), and $\eta^{\text{DPG},2}$ (—■—)

The Lamé parameter $\lambda = E\nu/((1+\nu)(1-2\nu))$ and $\mu = E/(2(1+\nu))$ with elastic modulus $E = 1$ and Poisson ratio $\nu = 4, 4.9, 4.99, 4.999, 4.9999$. Figure 5.10 shows that Poisson ratios ν close to 0.5 (and so large parameters λ) result in a pre-asymptotic regime without improvement of the error $\|\varepsilon(u_{\mathrm{ref}} - u_h)\|_{L^2(\Omega)}$ for the solution to the asymptotically exact DPG method $u_h = u_h^{\mathrm{DPG},2}$ and the Courant-FEM $u_h = u_h^{\mathrm{C}}$. Let $\mathrm{ndof} := \dim S^1(\mathcal{T};\mathbb{R}^d)$. The LSFEM and DPG method with operators (5.104) show (optimal) convergence

$$\|\varepsilon(u_{\mathrm{ref}} - u_h^{\mathrm{DPG},1})\|_{L^2(\Omega)} = \mathcal{O}(\mathrm{ndof}^{-0.5}) = \|\varepsilon(u_{\mathrm{ref}} - u_h^{\mathrm{LS}})\|_{L^2(\Omega)}$$

without pre-asymptotic regime, that is, the solutions do not lock. Moreover, the solutions $u_h^{\mathrm{DPG},1}$ and u_h^{LS} are almost identical. However, numerical difficulties (see Remark 5.2.14) cause the failure of the DPG method with operators (5.104), that is, for values ν close to 0.5 and small mesh-sizes the error $\|\varepsilon(u_{\mathrm{ref}} - u_h^{\mathrm{DPG},1})\|_{L^2(\Omega)}$ and the residual $\eta^{\mathrm{DPG},1}$ increase. Thus, the primal DPG method 1 behaves superior to the DPG method 2 and the Courant-FEM for Poisson ratios ν close to 0.5 and coarse meshes, but fails for fine meshes. This failure results from the large condition number of the Gram matrix G (and so a large numerical error in the inversion of G). Hence, the DPG method 1 is not competitive. This failure manifests a practical difficulty of the DPG method: The computation of the inverse Gram matrix G^{-1} (see Listing A.13 in Appendix A.2) leads to numerical difficulties for small mesh-sizes. A possible remedy are local mesh-dependent weights as for example in [CHBTD14] for convection-dominated diffusion problems. A first intuitive approach utilizes the local mesh-size $h \in \mathbb{P}_0(\mathcal{T})$ with $h|_T = \mathrm{diam}(T)$ for all $T \in \mathcal{T}$ as weight in the squared test norm

$$\|\bullet\|_{Y_w}^2 := \|h^{w/2}\mathbb{C}\nabla_{NC}\bullet\|_{L^2(\Omega)}^2 + \|\bullet\|_{L^2(\Omega)}^2 \qquad \text{for } w = 0, 1, 2$$

and computes the solution to the weighted DPG method with Poisson ratio $\nu = 0.49999$

$$\mathfrak{u}_h(w) = (u_h(w), s_h(w)) = \operatorname*{arg\,min}_{x_h \in X_h} \sup_{y_h \in Y_h \setminus \{0\}} \frac{b(x_h, y_h) - (f, y_h)_{L^2(\Omega)}}{\|y_h\|_{Y_w}} \qquad \text{for } w = 1, 2.$$

The left-hand side of Figure 5.11 plots the error $\|\varepsilon(u_{\mathrm{ref}} - u_h(w))\|_{L^2(\Omega)}$ with reference solution u_{ref}, adaptively refined meshes $\mathcal{T}$, and discrete solution $\mathfrak{u}_h(\mathrm{LS}) = (u_h(\mathrm{LS}), s_h(\mathrm{LS})) := \mathfrak{u}_h^{\mathrm{LS}}$, $\mathfrak{u}_h(0) = (u_h(0), s_h(0)) := \mathfrak{u}_h^{\mathrm{DPG},1}$ from the experiment in Figure 5.10. The right-hand side plots the residuals $\eta(\mathrm{LS}) := \eta^{\mathrm{LS}}$, $\eta(0) := \eta^{\mathrm{DPG},1}$, and

$$\eta(w) := \sup_{y_h \in Y_h \setminus \{0\}} \frac{b(\mathfrak{u}_h(w), y_h) - (f, y_h)_{L^2(\Omega)}}{\|y_h\|_{Y_w}} \qquad \text{for } w = 1, 2.$$

The weights $w = 1, 2$ extend the regime where the condition number of the Gram matrix G allows for accurate numerical inversion. However, the weight $w = 1$ decreases the rate of convergence and the error of the DPG method increases for triangulations with $3802 \leq \mathrm{ndof} = \dim S^1(\mathcal{T};\mathbb{R}^2)$. The weighted DPG method with weight $w = 2$ seems to be stable, that is, numerical difficulties do not result in large errors. Unfortunately, the error does not decrease as the mesh is refined. Thus, both weighed DPG methods fail and so the design of a numerical stable locking-free primal DPG method remains an open problem.

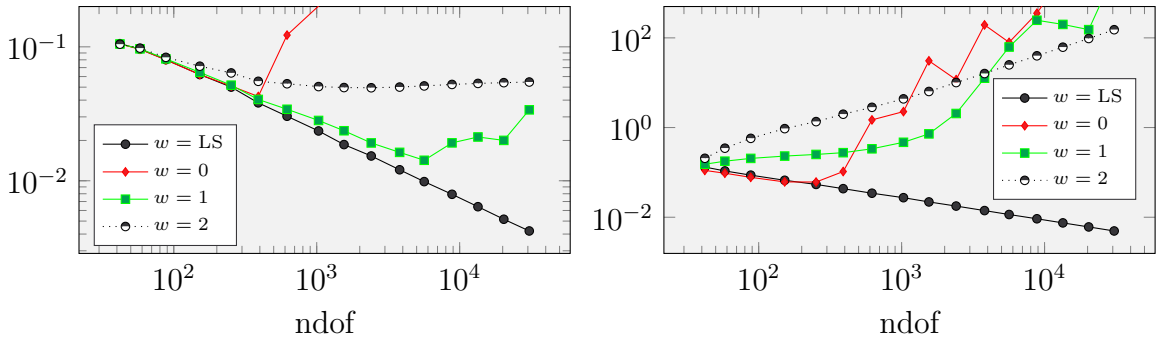


Figure 5.11: Convergence history plot of the error $\|\varepsilon(u_{\mathrm{ref}} - u_h(w))\|_{L^2(\omega)}$ (right) and residual $\eta(w)$ (left) with $w \in \{\mathrm{LS}, 0, 1, 2\}$ plotted agains ndof $= \dim S^1(\mathcal{T}; \mathbb{R}^2)$

6 Conclusion and outlook

Error control for minimal residual methods. This thesis introduces asymptotic exactness and best approximation results as well as an efficient guaranteed error control for linear model problems. Numerical experiments indicate an efficiency index close to one and so underline the advantages of the improved guaranteed error bound. A comparison of the natural built-in error control and the improved guaranteed error bound shows a significant improvement for the Helmholtz and Maxwell equations with frequency ω^2 close to an eigenvalue. This motivates the costly computation of the improved reliability constant. All proofs base on very general ideas, which might extend to further problems. These problems include LSFEMs for higher-order partial differential equations like the biharmonic equation [Tha00], modified formulations like the first-order system least squares methods for linear elasticity in [SSS10, SSS11], coupled LSFEMs like in [MS11], or non-linear LSFEMs like in [KMS17] for the Signorini problem and in [MS16, MSSS14, Sta07, Sta10, SSS09] for elasticity.

Moreover, the analysis and the numerical experiments indicate that eigenvalues close to the frequency ω^2 cause severe difficulties of the DPG method and the LSFEM for the Helmholtz and Maxwell equations. The design of a practical minimal residual method for problems with eigenvalues close to the frequency has to circumvent this difficulty. Maybe, an artificial shift of the spectrum allows for an improved numerical scheme.

Analysis of the DPG method. This thesis introduces an abstract framework for DPG methods which improves existing results. The analysis of the abstract framework suggests weighted test spaces for ultra-weak DPG methods. However, numerical results in [GMO14] indicate significant improvements for a weighted DPG method with weights which are not optimal in the context of the analysis in Section 5.1.6. This motivates further investigations. A related topic are local weights, which require a modification of the analysis in Section 5.1.5–5.1.6. Local weights might be beneficial for problems with varying material parameters. Moreover, mesh dependent weights might be necessary for the numerical stable inversion of the local Gram matrices in the computation of the discrete solution to the DPG method.

The abstract framework from Section 5.1.4 allows to design novel DPG methods for parabolic and hyperbolic problems. The conforming approximation of these traces is often unclear and so motivates non-conforming DPG methods. The analysis of these non-conforming schemes might utilize the idea of conforming reconstructions from [Ern18, EW19].

The analysis of ultra-weak DPG methods in Section 5.1.6 applies to triangulations with curved element and hanging nodes. The implementation of these schemes is simple. This motivates a comparison of ultra-weak DPG methods and finite element methods on curved boundaries like [BMS14a, BMS14b].

Computation of the LBB constant. This thesis utilizes the LSFEM to transform a challenging eigenvalue problem into an eigenvalue problem in a Rayleigh-Ritz-like environment. This approach leads to a convergent numerical scheme for the approximation of the LBB constant with beneficial properties. If the LBB constant belongs to an isolated eigenvalue, the experiments suggest that the adaptive scheme, driven by an heuristic error indicator, leads to optimal convergence and the error indicator is equivalent to the error. If the LBB constant belongs to the essential spectrum of the Cosserat operator, the convergence rate of the adaptive algorithm is poor. This shows similarities to adaptive finite element methods for eigenvalue problems with compact operators in [CG12, CGS15, DHZ15, DXZ08], which experience difficulties with eigenvalue clusters. The techniques from [BGGG17, Gal14b, Gal15] for adaptive schemes with clustered eigenvalues circumvent these difficulties and improve the convergence rates. This motivates the application of similar techniques to design adaptive schemes for the approximation of the LBB constant.

The transformation of challenging eigenvalue problems into an eigenvalue problem in a Rayleigh-Ritz-like environment with LSFEM might apply to more problems, for example non-symmetric eigenvalue problems. This motivates further investigations.

Bibliography

[ADM06] G. Acosta, R. G. Durán, and M.-A. Muschietti, *Solutions of the divergence operator on John domains*, Adv. Math. **206** (2006), 373–401.

[AQ92] D. N. Arnold and J. Qin, *Quadratic velocity/linear pressure Stokes elements*, Advances in computer methods for partial differential equations **7** (1992), 28–34.

[BA72] I. M. Babuška and A. K. Aziz, *Survey lectures on the mathematical foundations of the finite element method*, Academic Press, New York, 1972.

[BAA+18] S. Balay, S. Abhyankar, M. Adams, J. Brown, P. Brune, K. Buschelman, Dalcin L., V. Eijkhout, W. Gropp, D. Karpeyev, D. Kaushik, M. Knepley, D. May, L. Curfman McInnes, R. Mills, T. Munson, K. Rupp, P. Sanan, B. Smith, S. Zampini, H. Zhang, and H. Zhang, *Petsc users manual: Revision 3.10*, Tech. report, Argonne National Lab.(ANL), Argonne, IL (United States), 2018.

[Bab71] I. M. Babuška, *Error-bounds for finite element method*, Numer. Math. **16** (1970/71), 322–333.

[Bar15] S. Bartels, *Numerical methods for nonlinear partial differential equations*, Springer Ser. in Comput. Math., vol. 47, Springer, Cham, 2015.

[BBB14] G. R. Barrenechea, L. Boulton, and N. Boussaïd, *Finite element eigenvalue enclosures for the Maxwell operator*, SIAM J. Sci. Comput. **36** (2014), A2887–A2906.

[BBB17] G. R. Barrenechea, L. Boulton, and N. Boussaïd, *Local two-sided bounds for eigenvalues of self-adjoint operators*, Numer. Math. **135** (2017), 953–986.

[BBF13] D. Boffi, F. Brezzi, and M. Fortin, *Mixed finite element methods and applications*, Springer Ser. in Comput. Math., vol. 44, Springer, Heidelberg, 2013.

[BC01] A. Buffa and P. Ciarlet, Jr., *On traces for functional spaces related to Maxwell's equations. I. An integration by parts formula in Lipschitz polyhedra*, Math. Methods Appl. Sci. **24** (2001), 9–30.

[BC02] S. Bartels and C. Carstensen, *Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. Part II: Higher order FEM*, Math. Comp. **71** (2002), 971–994.

[BC04] ———, *Averaging techniques yield reliable a posteriori finite element error control for obstacle problems*, Numer. Math. **99** (2004), 225–249.

[BC05] C. Bahriawati and C. Carstensen, *Three Matlab implementations of the lowest-order Raviart-Thomas MFEM with a posteriori error control*, Comput. Methods Appl. Math. **5** (2005), 333–361.

[BC17a] P. Bringmann and C. Carstensen, *An adaptive least-squares FEM for the Stokes equations with optimal convergence rates*, Numer. Math. **135** (2017), 459–492.

[BC17b] _____, *h-adaptive least-squares finite element methods for the 2D Stokes equations of any order with optimal convergence rates*, Comput. Math. Appl. **74** (2017), 1923–1939.

[BCCO09] S. Bond, J. Chaudhry, E. Cyr, and L. Olson, *A first-order system least-squares finite element method for the Poisson-Boltzmann equation*, J. Comput. Chem. **31** (2009), 1625–35.

[BCDG16] C. Bernardi, M. Costabel, M. Dauge, and V. Girault, *Continuity properties of the inf-sup constant for the divergence*, SIAM J. Math. Anal. **48** (2016), 1250–1271.

[BCM16] P. Bringmann, C. Carstensen, and C. Merdon, *Guaranteed velocity error control for the pseudostress approximation of the Stokes equations*, Numer. Methods Partial Differential Equations **32** (2016), 1411–1432.

[BCS02] A. Buffa, M. Costabel, and D. Sheen, *On traces for $H(\mathrm{curl}, \Omega)$ in Lipschitz domains*, J. Math. Anal. Appl. **276** (2002), 845–867.

[BCS18] P. Bringmann, C. Carstensen, and G. Starke, *An adaptive least-squares FEM for linear elasticity with optimal convergence rates*, SIAM J. Numer. Anal. **56** (2018), 428–447.

[BDD04] P. Binev, W. Dahmen, and R. DeVore, *Adaptive finite element methods with convergence rates*, Numer. Math. **97** (2004), 219–268.

[BDS18] D. Broersen, W. Dahmen, and R. P. Stevenson, *On the stability of DPG formulations of transport equations*, Math. Comp. **87** (2018), 1051–1082.

[Bey95] J. Bey, *Tetrahedral grid refinement*, Computing **55** (1995), 355–378.

[BG94] P. B. Bochev and M. D. Gunzburger, *Analysis of least squares finite element methods for the Stokes equations*, Math. Comp. **63** (1994), 479–506.

[BG98] _____, *Finite element methods of least-squares type*, SIAM Rev. **40** (1998), 789–837.

[BG05] _____, *On least-squares finite element methods for the Poisson equation and their connection to the Dirichlet and Kelvin principles*, SIAM J. Numer. Anal. **43** (2005), 340–362.

[BG09] _____, *Least-squares finite element methods*, Appl. Math. Sci., vol. 166, Springer, New York, 2009.

[BGGG17] D. Boffi, D. Gallistl, F. Gardini, and L. Gastaldi, *Optimal convergence of adaptive FEM for eigenvalue clusters in mixed form*, Math. Comp. **86** (2017), 2213–2237.

[BHHW00] R. Beck, R. Hiptmair, R. H. W. Hoppe, and B. Wohlmuth, *Residual based a posteriori error estimators for eddy current computation*, M2AN Math. Model. Numer. Anal. **34** (2000), 159–182.

[BHP17] A. Bespalov, A. Haberl, and D. Praetorius, *Adaptive FEM with coarse initial mesh guarantees optimal convergence rates for compactly perturbed elliptic problems*, Comput. Methods Appl. Mech. Engrg. **317** (2017), 318–340.

[BKP05a] J. H. Bramble, T. V. Kolev, and J. E. Pasciak, *The approximation of the Maxwell eigenvalue problem using a least-squares method*, Math. Comp. **74** (2005), 1575–1598.

[BKP05b] ———, *A least-squares approximation method for the time-harmonic Maxwell equations*, J. Numer. Math. **13** (2005), 237–263.

[BM08] R. Becker and S. Mao, *An optimally convergent adaptive mixed finite element method*, Numer. Math. **111** (2008), 35–54.

[BMM97] M. Berndt, T. A. Manteuffel, and S. F. McCormick, *Local error estimates and adaptive refinement for first-order system least squares (FOSLS)*, Electron. Trans. Numer. Anal. **6** (1997), 35–43, Special issue on multilevel methods (Copper Mountain, CO, 1997).

[BMS] F. Bertrand, M. Moldenhauer, and G. Starke, *A posteriori error estimation for planar linear elasticity by stress reconstruction*, to appear in Comput. Methods Appl. Math.

[BMS14a] F. Bertrand, S. Münzenmaier, and G. Starke, *First-order system least squares on curved boundaries: higher-order Raviart-Thomas elements*, SIAM J. Numer. Anal. **52** (2014), 3165–3180.

[BMS14b] ———, *First-order system least squares on curved boundaries: lowest-order Raviart-Thomas elements*, SIAM J. Numer. Anal. **52** (2014), 880–894.

[Bog79] M. E. Bogovskiĭ, *Solution of the first boundary value problem for an equation of continuity of an incompressible medium*, Soviet. Math. Doklady **20** (1979), 1094–1098.

[Bra07] D. Braess, *Finite elements: Theory, fast solvers, and applications in elasticity theory*, third ed., Cambridge University Press, Cambridge, 2007.

[Bre74] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge **8** (1974), 129–151.

[BS92a] I. M. Babuška and M. Suri, *Locking effects in the finite element approximation of elasticity problems*, Numer. Math. **62** (1992), 439–463.

[BS92b] ———, *On locking and robustness in the finite element method*, SIAM J. Numer. Anal. **29** (1992), 1261–1293.

[BS97] I. M. Babuška and S. A. Sauter, *Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers?*, SIAM J. Numer. Anal. **34** (1997), 2392–2423.

[BS02] S. C. Brenner and L. R. Scott, *The mathematical theory of finite element methods*, second ed., Texts Appl. Math., vol. 15, Springer, New York, 2002.

[BS08] D. Braess and J. Schöberl, *Equilibrated residual error estimator for edge elements*, Math. Comp. **77** (2008), 651–672.

[BV00] C. Bernardi and R. Verfürth, *Adaptive finite element methods for elliptic equations with non-smooth coefficients*, Numer. Math. **85** (2000), 579–608.

[CA03] C. Carstensen and J. Alberty, *Averaging techniques for reliable a posteriori FE-error control in elastoplasticity with hardening*, Comput. Methods Appl. Mech. Engrg. **192** (2003), 1435–1450.

[Car99] C. Carstensen, *Quasi-interpolation and a posteriori error analysis in finite element methods*, M2AN Math. Model. Numer. Anal. **33** (1999), 1187–1202.

[Car02] ———, *Merging the Bramble-Pasciak-Steinbach and the Crouzeix-Thomée criterion for H^1-stability of the L^2-projection onto finite element spaces*, Math. Comp. **71** (2002), 157–163.

[Car04] ———, *All first-order averaging techniques for a posteriori finite element error control on unstructured grids are efficient and reliable*, Math. Comp. **73** (2004), 1153–1165.

[CB02] C. Carstensen and S. Bartels, *Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids part I: Low order conforming, nonconforming, and mixed FEM*, Math. Comp. **71** (2002), 945–969.

[CBHW18] C. Carstensen, P. Bringmann, F. Hellwig, and P. Wriggers, *Nonlinear discontinuous Petrov–Galerkin methods*, Numer. Math. **139** (2018), 529–561.

[CCDL15] M. Costabel, M. Crouzeix, M. Dauge, and Y. Lafranche, *The inf-sup constant for the divergence on corner domains*, Numer. Methods Partial Differential Equations **31** (2015), 439–458.

[CCKP15] Z. Cai, V. Carey, J. Ku, and E.-J. Park, *Asymptotically exact a posteriori error estimators for first-order div least-squares methods in local and global L_2 norm*, Comput. Math. Appl. **70** (2015), 648–659.

[CD98] C. Carstensen and G. Dolzmann, *A posteriori error estimates for mixed FEM in elasticity*, Numer. Math. **81** (1998), 187–209.

[CD15] M. Costabel and M. Dauge, *On the inequalities of Babuška–Aziz, Friedrichs and Horgan–Payne*, Arch. Ration. Mech. Anal. **217** (2015), 873–898.

[CDFH00] C. Carstensen, G. Dolzmann, S. A. Funken, and D. S. Helm, *Locking-free adaptive mixed finite element methods in linear elasticity*, Comput. Methods Appl. Mech. Engrg. **190** (2000), 1701–1718.

[CDG14] C. Carstensen, L. Demkowicz, and J. Gopalakrishnan, *A posteriori error control for DPG methods*, SIAM J. Numer. Anal. **52** (2014), 1335–1353.

[CDG16] ———, *Breaking spaces and forms for the DPG method and applications including Maxwell equations*, Comput. Math. Appl. **72** (2016), 494–522.

[CDW12] A. Cohen, W. Dahmen, and G. Welper, *Adaptivity and variational stabilization for convection-diffusion equations*, ESAIM Math. Model. Numer. Anal. **46** (2012), 1247–1273.

[CF99] C. Carstensen and S. A. Funken, *Fully reliable localized error control in the FEM*, SIAM J. Sci. Comput. **21** (1999), 1465–1484 (electronic).

[CF01a] ———, *Averaging technique for a posteriori error control in elasticity. III. Locking-free nonconforming FEM*, Comput. Methods Appl. Mech. Engrg. **191** (2001), 861–877.

[CF01b] ———, *Averaging technique for FE-a posteriori error control in elasticity. part I: Conforming fem*, Comput. Methods Appl. Mech. Engrg. **190** (2001), 2483–2498.

[CF01c] ———, *Averaging technique for FE-a posteriori error control in elasticity. part II: λ-independent estimates*, Comput. Methods Appl. Mech. Engrg. **190** (2001), 4663–4675.

[CF00] ———, *Fully reliable localized error control in the FEM*, SIAM J. Sci. Comput. **21** (1999/00), 1465–1484.

[CFPP14] C. Carstensen, M. Feischl, M. Page, and D. Praetorius, *Axioms of adaptivity*, Comput. Math. Appl. **67** (2014), 1195–1253.

[CG12] C. Carstensen and J. Gedicke, *An adaptive finite element eigenvalue solver of asymptotic quasi-optimal computational complexity*, SIAM J. Numer. Anal. **50** (2012), 1029–1057.

[CG14a] C. Carstensen and D. Gallistl, *Guaranteed lower eigenvalue bounds for the biharmonic equation*, Numer. Math. **126** (2014), 33–51.

[CG14b] C. Carstensen and J. Gedicke, *Guaranteed lower bounds for eigenvalues*, Math. Comp. **83** (2014), 2605–2629.

[CGHW14] C. Carstensen, D. Gallistl, F. Hellwig, and L. Weggler, *Low-order dPG-FEM for an elliptic PDE*, Comput. Math. Appl. **68** (2014), 1503–1512.

[CGS13a] C. Carstensen, D. Gallistl, and M. Schedensack, *Discrete reliability for Crouzeix-Raviart FEMs*, SIAM J. Numer. Anal. **51** (2013), 2935–2955.

[CGS13b] ———, *Quasi-optimal adaptive pseudostress approximation of the Stokes equations*, SIAM J. Numer. Anal. **51** (2013), 1715–1734.

[CGS15] ———, *Adaptive nonconforming Crouzeix-Raviart FEM for eigenvalue problems*, Math. Comp. **84** (2015), 1061–1087.

[CH16] C. Carstensen and F. Hellwig, *Low-order discontinuous Petrov-Galerkin finite element methods for linear elasticity*, SIAM J. Numer. Anal. **54** (2016), 3388–3410.

[CH18] ———, *Optimal Convergence Rates for Adaptive Lowest-Order Discontinuous Petrov-Galerkin Schemes*, SIAM J. Numer. Anal. **56** (2018), 1091–1111.

[CHBTD14] J. Chan, N. Heuer, T. Bui-Thanh, and L. Demkowicz, *A robust DPG method for convection-dominated diffusion problems II: adjoint boundary conditions and mesh-dependent test norms*, Comput. Math. Appl. **67** (2014), 771–795.

[Cia78] P. G. Ciarlet, *The finite element method for elliptic problems*, vol. 4, North-Holland, Amsterdam, 1978.

[CKNS08] J. M. Cascón, C. Kreuzer, R. H. Nochetto, and K. G. Siebert, *Quasi-optimal convergence rate for an adaptive finite element method*, SIAM J. Numer. Anal. **46** (2008), 2524–2550.

[CKP11] C. Carstensen, D. Kim, and E.-J. Park, *A priori and a posteriori pseudostress-velocity mixed finite element error analysis for the Stokes problem*, SIAM J. Numer. Anal. **49** (2011), 2501–2523.

[CKS05] Z. Cai, J. Korsawe, and G. Starke, *An adaptive least squares mixed finite element method for the stress-displacement formulation of linear elasticity*, Numer. Methods Partial Differential Equations **21** (2005), 132–148.

[CL91] P. G. Ciarlet and J.-L. Lions (eds.), *Handbook of numerical analysis. Vol. II. Finite element methods. Part 1*, North-Holland, Amsterdam, 1991.

[CLMM94] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part I*, SIAM J. Numer. Anal. **31** (1994), 1785–1799.

[CLW04] Z. Cai, B. Lee, and P. Wang, *Least-squares methods for incompressible Newtonian fluid flow: Linear stationary problems*, SIAM J. Numer. Anal. **42** (2004), 843–859.

[CM10] C. Carstensen and C. Merdon, *Estimator competition for Poisson problems*, J. Comput. Math. **28** (2010), 309–330.

[CM11] ———, *Remarks on the state of the art of a posteriori error control of elliptic PDEs in energy norms in practise*, Stud. Univ. Babeş-Bolyai Math. **56** (2011), 273–293.

[CM13] ———, *Effective postprocessing for equilibration a posteriori error estimators*, Numer. Math. **123** (2013), 425–459.

[CM14] ———, *Refined fully explicit a posteriori residual-based error control*, SIAM J. Numer. Anal. **52** (2014), 1709–1728.

[CN12] J. M. Cascón and R. H. Nochetto, *Quasioptimal cardinality of AFEM driven by nonresidual estimators*, IMA J. Numer. Anal. **32** (2012), 1–29.

[CP15] C. Carstensen and E.-J. Park, *Convergence and optimality of adaptive least squares finite element methods*, SIAM J. Numer. Anal. **53** (2015), 43–62.

[CP18] C. Carstensen and S. Puttkammer, *A low-order discontinuous Petrov–Galerkin method for the Stokes equations*, Numer. Math. **140** (2018), 1–34.

[CPB17] C. Carstensen, E.-J. Park, and P. Bringmann, *Convergence of natural adaptive least squares FEMs*, Numer. Math. **136** (2017), 1097–1115.

[CR11] C. Carstensen and H. Rabus, *An optimal adaptive mixed finite element method*, Math. Comp. **80** (2011), 649–667.

[CR17] ———, *Axioms of adaptivity with separate marking for data resolution*, SIAM J. Numer. Anal. **55** (2017), 2644–2665.

[Cro97] M. Crouzeix, *On an operator related to the convergence of Uzawa's algorithm for the Stokes equation*, Computational science for the 21st century, Wiley, 1997, pp. 242–249.

[CS03] Z. Cai and G. Starke, *First-order system least squares for the stress-displacement formulation: linear elasticity*, SIAM J. Numer. Anal. **41** (2003), 715–730.

[CS04] ———, *Least-squares methods for linear elasticity*, SIAM J. Numer. Anal. **42** (2004), 826–842.

[CS18] C. Carstensen and J. Storn, *Asymptotic exactness of the least-squares finite element residual*, SIAM J. Numer. Anal. **56** (2018), 2008–2028.

[CTVW10] Z. Cai, C. Tong, P. S. Vassilevski, and C. Wang, *Mixed finite element methods for incompressible flow: Stationary Stokes equations*, Numer. Methods Partial Differential Equations **26** (2010), 957–978.

[CV99] C. Carstensen and R. Verfürth, *Edge residuals dominate a posteriori error estimates for low order finite element methods*, SIAM J. Numer. Anal. **36** (1999), 1571–1587.

[CW07] Z. Cai and Y. Wang, *A multigrid method for the pseudostress formulation of Stokes problems*, SIAM J. Sci. Comput. **29** (2007), 2078–2095.

[CW10] _____, *Pseudostress-velocity formulation for incompressible Navier-Stokes equations*, Internat. J. Numer. Methods Fluids **63** (2010), 341–356.

[CWZ10] Z. Cai, C. Wang, and S. Zhang, *Mixed finite element methods for incompressible flow: stationary Navier-Stokes equations*, SIAM J. Numer. Anal. **48** (2010), 79–94.

[CY00] Z. Cai and X. Ye, *A least-squares finite element approximation for the compressible Stokes equations*, Numer. Methods Partial Differential Equations **16** (2000), 62–70.

[Dau88] M. Dauge, *Elliptic boundary value problems on corner domains. Smoothness and asymptotics of solutions*, Lecture Notes in Math., vol. 1341, Springer, Berlin, 1988.

[DG10] L. Demkowicz and J. Gopalakrishnan, *A class of discontinuous Petrov-Galerkin methods. Part I: The transport equation*, Comput. Methods Appl. Mech. Engrg. **199** (2010), 1558–1572.

[DG11] _____, *A class of discontinuous Petrov-Galerkin methods. Part II: Optimal test functions*, Numer. Methods Partial Differential Equations **27** (2011), 70–105.

[DGN12] L. Demkowicz, J. Gopalakrishnan, and A. H. Niemi, *A class of discontinuous Petrov-Galerkin methods. Part III: Adaptivity*, Appl. Numer. Math. **62** (2012), 396–427.

[DGNS17] L. Demkowicz, J. Gopalakrishnan, S. Nagaraj, and P. Sepúlveda, *A spacetime DPG method for the Schrödinger equation*, SIAM J. Numer. Anal. **55** (2017), 1740–1759.

[DHZ15] X. Dai, L. He, and A. Zhou, *Convergence and quasi-optimal complexity of adaptive finite element computations for multiple eigenvalues*, IMA J. Numer. Anal. **35** (2015), 1934–1977.

[DKS13] L. Diening, C. Kreuzer, and E. Süli, *Finite element approximation of steady flows of incompressible fluids with implicit power-law-like rheology*, SIAM J. Numer. Anal. **51** (2013), 984–1015.

[DMRT10] R. G. Durán, M.-A. Muschietti, E. Russ, and P. Tchamitchian, *Divergence operator and Poincaré inequalities on arbitrary bounded domains*, Complex Var. Elliptic Equ. **55** (2010), 795–816.

[DPKC11] L. D. Dalcin, R. R. Paz, P. A. Kler, and A. Cosimo, *Parallel distributed computing using python*, Adv. Water Resources **34** (2011), 1124–1139, New Computational Methods and Software Tools.

[Dur12] R. G. Durán, *An elementary proof of the continuity from $L^2_0(\Omega)$ to $H^1_0(\Omega)^n$ of Bogovskii's right inverse of the divergence*, Rev. Un. Mat. Argentina **53** (2012), 59–78.

[DV98] L. Demkowicz and L. Vardapetyan, *Modeling of electromagnetic absorption/ scattering problems using hp-adaptive finite elements*, Comput. Methods Appl. Mech. Engrg. **152** (1998), 103–124.

[DXZ08] X. Dai, J. Xu, and A. Zhou, *Convergence and optimal complexity of adaptive finite element eigenvalue computations*, Numer. Math. **110** (2008), 313–355.

[Ern18] J. Ernesti, *Space-time methods for acoustic waves with applications to full waveform inversion*, Doctoral dissertation, Karlsruher Institut für Technologie, 2018.

[EV15] A. Ern and M. Vohralík, *Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations*, SIAM J. Numer. Anal. **53** (2015), 1058–1081.

[Eva10] L. C. Evans, *Partial differential equations*, second ed., Grad. Stud. Math., vol. 19, Amer. Math. Soc., Providence, RI, 2010.

[EW19] J. Ernesti and C. Wieners, *A space-time discontinuous Petrov-Galerkin method for acoustic waves*, Space-Time Methods. Applications to Partial Differential Equations, Radon Ser. Comput. Appl. Math., De Gruyter, Berlin, 2019, to appear.

[FFP14] M. Feischl, T. Führer, and D. Praetorius, *Adaptive FEM with optimal convergence rates for a certain class of nonsymmetric and possibly nonlinear problems*, SIAM J. Numer. Anal. **52** (2014), 601–625.

[FGM09] L. E. Figueroa, G. N. Gatica, and A. Márquez, *Augmented mixed finite element methods for the stationary Stokes equations*, SIAM J. Sci. Comput. **31** (2008/09), 1082–1119.

[FHS18] T. Führer, N. Heuer, and E. P. Stephan, *On the DPG method for Signorini problems*, IMA Journal of Numerical Analysis **38** (2018), 1893–1926.

[FHSG17] T. Führer, N. Heuer, and J. Sen Gupta, *A time-stepping DPG scheme for the heat equation*, Comput. Methods Appl. Math. **17** (2017), 237–252.

[FMM98] J. M. Fiard, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares (FOSLS) for convection-diffusion problems: numerical results*, SIAM J. Sci. Comput. **19** (1998), 1958–1979.

[Fri37] K. O. Friedrichs, *On certain inequalities and characteristic value problems for analytic functions and for functions of two variables*, Trans. Amer. Math. Soc. **41** (1937), 321–364.

[Fri47] ———, *On the boundary-value problems of the theory of elasticity and Korn's inequality*, Ann. of Math. **48** (1947), 441–471.

[Gal14a] D. Gallistl, *Adaptive finite element computation of eigenvalues*, Doctoral dissertation, Humboldt-Universität zu Berlin, 2014.

[Gal14b] ———, *Adaptive nonconforming finite element approximation of eigenvalue clusters*, Comput. Methods Appl. Math. **14** (2014), 509–535.

[Gal15] ———, *An optimal adaptive FEM for eigenvalue clusters*, Numer. Math. **130** (2015), 467–496.

[Gal19] ———, *Rayleigh-Ritz approximation of the inf-sup constant for the divergence*, Math. Comp. **88** (2019), 73–89.

[Gel81] P. B. Geltner, *General Rayleigh quotient iteration*, SIAM J. Numer. Anal. **18** (1981), 839–843.

[GHS16] F. D. Gaspoz, C.-J. Heine, and K. G. Siebert, *Optimal grading of the newest vertex bisection and H^1-stability of the L_2-projection*, IMA J. Numer. Anal. **36** (2016), 1217–1241.

[GMO14] J. Gopalakrishnan, I. Muga, and N. Olivares, *Dispersive and dissipative errors in the DPG method with scaled norms for Helmholtz equation*, SIAM J. Sci. Comput. **36** (2014), 20–39.

[GMS10] G. N. Gatica, A. Márquez, and M. A. Sánchez, *Analysis of a velocity-pressure-pseudostress formulation for the stationary Stokes equations*, Comput. Methods Appl. Mech. Engrg. **199** (2010), 1064–1079.

[GMS11] ———, *A priori and a posteriori error analyses of a velocity-pseudostress formulation for a class of quasi-Newtonian Stokes flows*, Comput. Methods Appl. Mech. Engrg. **200** (2011), 1619–1636.

[GN14a] J. Guzmán and M. Neilan, *Conforming and divergence-free Stokes elements in three dimensions*, IMA J. Numer. Anal. **34** (2014), 1489–1508.

[GN14b] ———, *Conforming and divergence-free Stokes elements on general triangular meshes*, Math. Comp. **83** (2014), 15–36.

[GQ14] J. Gopalakrishnan and W. Qiu, *An analysis of the practical DPG method*, Math. Comp. **83** (2014), 537–552.

[GR86] V. Girault and P.-A. Raviart, *Finite element methods for Navier-Stokes equations: Theory and algorithms*, Springer Ser. Comput. Math., vol. 5, Springer, Berlin, 1986.

[GS17] J. Gopalakrishnan and P. Sepulveda, *A spacetime DPG method for acoustic waves*, arXiv preprint arXiv:1709.08268 (2017).

[GSS14] D. Gallistl, M. Schedensack, and R. P. Stevenson, *A remark on newest vertex bisection in any space dimension*, Comput. Methods Appl. Math. **14** (2014), 317–320.

[Hel18] F. Hellwig, *Adaptive discontinuous Petrov-Galerkin Finite-Element-Methods*, Doctoral dissertation, Humboldt-Universität zu Berlin, 2018.

[HHNL88] I. Hlaváček, J. Haslinger, J. Nečas, and J. Lovíšek, *Solution of variational inequalities in mechanics*, Applied Mathematical Sciences, vol. 66, Springer-Verlag, New York, 1988.

[HL11] Q. Han and F. Lin, *Elliptic partial differential equations*, second ed., Courant Lect. Notes Math., vol. 1, Courant Inst. Math. Sci., New York, 2011.

[HP83] C. O. Horgan and L. E. Payne, *On inequalities of Korn, Friedrichs and Babuška-Aziz*, Arch. Rational Mech. Anal. **82** (1983), 165–179.

[HRTV07] V. Hernández, J. E. Román, A. Tomás, and V. Vidal, *Krylov-schur methods in SLEPc*, Universitat Politecnica de Valencia, Tech. Rep. STR-7 (2007), 1–13.

[Ily09] A. A. Ilyin, *On the spectrum of the Stokes operator*, Funktsional. Anal. i Prilozhen. **43** (2009), 14–25.

[JP93] B.-N. Jiang and L. A. Povinelli, *Optimal least-squares finite element method for elliptic problems*, Comput. Methods Appl. Mech. Engrg. **102** (1993), 199–212.

[KKR+17] B. Keith, P. Knechtges, N. V. Roberts, S. Elgeti, M. Behr, and L. Demkowicz, *An ultraweak DPG method for viscoelastic fluids*, J. Non-Newton. Fluid Mech. **247** (2017), 107–122.

[KMS17] R. Krause, B. Müller, and G. Starke, *An adaptive least-squares mixed finite element method for the Signorini problem*, Numer. Methods Partial Differential Equations **33** (2017), 276–289.

[KVZ+72] M. A. Krasnosel'skii, G. M. Vainikko, P. P. Zabreiko, Y. B. Rutitskii, and V. Y. Stetsenko, *Approximate solution of operator equations*, Noordhoff, Groningen, 1972.

[Lad63] O. A. Ladyzhenskaya, *The mathematical theory of viscous incompressible flow*, Revised English edition., Gordon and Breach Science Publishers, New York-London, 1963.

[Lin07] A. Linke, *Divergence-free mixed finite elements for the incompressible Navier-Stokes equation*, Doctoral dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2007.

[LMW12] A. Logg, K.-A. Mardal, and G. N. Wells (eds.), *Automated solution of differential equations by the finite element method: The FEniCS book*, Lect. Notes Comput. Sci. Eng., vol. 84, Springer, Heidelberg, 2012.

[Mau95] J. M. Maubach, *Local bisection refinement for n-simplicial grids generated by reflection*, SIAM J. Sci. Comput. **16** (1995), 210–227.

[McL00] W. McLean, *Strongly elliptic systems and boundary integral equations*, Cambridge University Press, Cambridge, 2000.

[Mit17] W. F. Mitchell, *30 years of newest vertex bisection*, JNAIAM. J. Numer. Anal. Ind. Appl. Math. **11** (2017), 11–22.

[MMR15] S. Meddahi, D. Mora, and R. Rodríguez, *A finite element analysis of a pseudostress formulation for the Stokes eigenvalue problem*, IMA J. Numer. Anal. **35** (2015), 749–766.

[Mon03] P. Monk, *Finite element methods for Maxwell's equations*, Numer. Math. Sci. Comput., Oxford Univ. Press, New York, 2003.

[MORR81] B. Mercier, J. Osborn, J. Rappaz, and P.-A. Raviart, *Eigenvalue approximation by mixed and hybrid methods*, Math. Comp. **36** (1981), 427–453.

[MS11] S. Münzenmaier and G. Starke, *First-order system least squares for coupled Stokes-Darcy flow*, SIAM J. Numer. Anal. **49** (2011), 387–404.

[MS16] B. Müller and G. Starke, *Stress-based finite element methods in linear and nonlinear solid mechanics*, pp. 69–104, Springer, Cham, 2016.

[MSSS14] B. Müller, G. Starke, A. Schwarz, and J. Schröder, *A first-order system least squares method for hyperelasticity*, SIAM J. Sci. Comput. **36** (2014), B795–B816.

[MW01] J. M. Melenk and B. I. Wohlmuth, *On residual-based a posteriori error estimation in hp-FEM*, Adv. Comput. Math. **15** (2001), 311–331.

[NS16] M. Neilan and D. Sap, *Stokes elements on cubic meshes yielding divergence-free approximations*, Calcolo **53** (2016), 263–283.

[PC94] A. I. Pehlivanov and G. F. Carey, *Error estimates for least-squares mixed finite elements*, RAIRO Modél. Math. Anal. Numér. **28** (1994), 499–516.

[PCL94] A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov, *Least-squares mixed finite elements for second-order elliptic problems*, SIAM J. Numer. Anal. **31** (1994), 1368–1377.

[PD17] S. Petrides and L. F. Demkowicz, *An adaptive DPG method for high frequency time-harmonic wave propagation problems*, Comput. Math. Appl. **74** (2017), 1999–2017.

[PSD09] N. Parés, H. Santos, and P. Díez, *Guaranteed energy error bounds for the Poisson equation using a flux-free approach: solving the local problems in subdomains*, Internat. J. Numer. Methods Engrg. **79** (2009), 1203–1244.

[PW60] L. E. Payne and H. F. Weinberger, *An optimal Poincaré inequality for convex domains*, Arch. Rational Mech. Anal. **5** (1960), 286–292.

[QZ07] J. Qin and S. Zhang, *Stability and approximability of the $\mathcal{P}1$–$\mathcal{P}0$ element for Stokes equations*, Internat. J. Numer. Methods Fluids **54** (2007), 497–515.

[RBTD14] N. V. Roberts, T. Bui-Thanh, and L. Demkowicz, *The DPG method for the Stokes problem*, Comput. Math. Appl. **67** (2014), 966–995.

[RDM15] N. V. Roberts, L. Demkowicz, and R. Moser, *A discontinuous Petrov-Galerkin methodology for adaptive solutions to the incompressible Navier-Stokes equations*, J. Comput. Phys. **301** (2015), 456–483.

[Rep97] S. I. Repin, *A posteriori error estimation for nonlinear variational problems by duality theory*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) **243** (1997), 201–214, 342.

[Rep98] ———, *A posteriori error estimation for approximate solutions of variational problems by duality theory*, ENUMATH 97 (Heidelberg), World Sci. Publ., River Edge, NJ, 1998, pp. 524–531.

[Rep99a] ———, *A posteriori estimates for the Stokes problem*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) **259** (1999), 195–211.

[Rep99b] ———, *A unified approach to a posteriori error estimation based on duality error majorants*, Math. Comput. Simulation **50** (1999), 305–321, Modelling '98 (Prague).

[Rep00] ———, *A posteriori error estimates for approximate solutions of variational problems with functionals of power growth*, J. Math. Sci. **101** (2000), 3531–3538.

[Rod94] R. Rodríguez, *Some remarks on Zienkiewicz-Zhu estimator*, Numer. Methods Partial Differential Equations **10** (1994), 625–635.

[RSS03] S. Repin, S. Sauter, and A. Smolianski, *A posteriori error estimation for the Dirichlet problem with account of the error in the approximation of boundary conditions*, Computing **70** (2003), 205–233.

[RX96] S. I. Repin and L. S. Xanthis, *A posteriori error estimation for elasto-plastic problems based on duality theory*, Comput. Methods Appl. Mech. Engrg. **138** (1996), 317–339.

[SH98] J. C. Simo and T. J. R. Hughes, *Computational inelasticity*, Interdiscip. Appl. Math., vol. 7, Springer, New York, 1998.

[SS05] E. M. Stein and R. Shakarchi, *Real analysis: Measure theory, integration, and Hilbert spaces*, Princeton Lect. Anal., vol. 3, Princeton Univ. Press, Princeton, NJ, 2005.

[SSS09] A. Schwarz, J. Schröder, and G. Starke, *Least-squares mixed finite elements for small strain elasto-viscoplasticity*, Internat. J. Numer. Methods Engrg. **77** (2009), 1351–1370.

[SSS10] A. Schwarz, J. Schröder, and G. Starke, *A modified least-squares mixed finite element with improved momentum balance*, Internat. J. Numer. Methods Engrg. **81** (2010), 286–306.

[SSS11] G. Starke, A. Schwarz, and J. Schröder, *Analysis of a modified first-order system least squares method for linear elasticity with improved momentum balance*, SIAM J. Numer. Anal. **49** (2011), 1006–1022.

[Sta07] G. Starke, *An adaptive least-squares mixed finite element method for elasto-plasticity*, SIAM J. Numer. Anal. **45** (2007), 371–388.

[Sta10] ———, *Adaptive least squares finite element methods in elasto-plasticity*, Large-Scale Scientific Computing, Springer, 2010, pp. 671–678.

[Ste07] R. Stevenson, *Optimality of a standard adaptive finite element method*, Found. Comput. Math. **7** (2007), 245–269.

[Ste08] ———, *The completion of locally refined simplicial partitions created by bisection*, Math. Comp. **77** (2008), 227–241.

[SV85] L. R. Scott and M. Vogelius, *Norm estimates for a maximal right inverse of the divergence operator in spaces of piecewise polynomials*, RAIRO Modél. Math. Anal. Numér. **19** (1985), 111–143.

[SY97] P. Shi and X. Ye, *A least-square mixed method for Stokes equations*, Numer. Methods Partial Differential Equations **13** (1997), 191–199.

[TB06] L. N. Trefethen and T. Betcke, *Computed eigenmodes of planar regions*, Recent advances in differential equations and mathematical physics, Contemp. Math., vol. 412, Amer. Math. Soc., Providence, RI, 2006, pp. 297–314.

[Tha00] R. W. Thatcher, *A least squares method for solving biharmonic problems*, SIAM J. Numer. Anal. **38** (2000), 1523–1539.

[Tra97] C. T. Traxler, *An algorithm for adaptive mesh refinement in n dimensions*, Computing **59** (1997), 115–137.

[Vel96] W. Velte, *On inequalities of Friedrichs and Babuška-Aziz*, Meccanica **31** (1996), 589–596.

[Vel98] ———, *On inequalities of Friedrichs and Babuška-Aziz in dimension three*, Z. Anal. Anwendungen **17** (1998), 843–857.

[Zha08] S. Zhang, *On the P1 Powell-Sabin divergence-free finite element for the Stokes equations*, J. Comput. Math. **26** (2008), 456–470.

[ZMD+11] J. Zitelli, I. Muga, L. Demkowicz, J. Gopalakrishnan, D. Pardo, and V. M. Calo, *A class of discontinuous Petrov-Galerkin methods. Part IV: The optimal test norm and time-harmonic wave propagation in 1D*, J. Comput. Phys. **230** (2011), 2406–2432.

[ZZ87] O. C. Zienkiewicz and J. Z. Zhu, *A simple error estimator and adaptive procedure for practical engineering analysis*, Internat. J. Numer. Methods Engrg. **24** (1987), 337–357.

Appendix

This appendix describes the implementation of the numerical experiments in this thesis. Appendix A.1 explains the routines for the LSFEM, Appendix A.2 explains the routines for the DPG method, and Appendix A.3.2 explains the adaptive mesh refinement and the computation of lower eigenvalue bounds with the Crouzeix-Raviart FEM. The digital version of this thesis contains all implemented routines as embedded zip-file. Appendix A.4 visualizes the directory structure of the zip-file.

A.1 Implementation of LSFEM

A.1.1 Computation of the solution

This section exemplifies the computation of the solution $\mathfrak{u}_h \in X_h$ to the LSFEMs from Chapter 3 with FEniCS 2017.2.0 by the function PMPLSFEM in SolverLSFEM.py. The function PMPLSFEM solves the LSFEM for the Poisson model problem from Section 3.2.1. Its structure reads:

1. define the function space and bilinear form,
2. set the boundary condition,
3. assemble the system matrices,
4. solve the linear system of equations,
5. compute the residual.

Input parameters are a regular triangulation $\mathcal{T}$, called mesh (an element in the class dolfin.cpp.mesh.Mesh()), the polynomial degree $k \in \mathbb{N}$ of the space $X_h = S_0^k(\mathcal{T}) \times RT_{k-1}(\mathcal{T})$ from (3.40), called polydegree, and the right-hand side $f \in L^2(\Omega)$, called f (an element in the class dolfin.functions.expression.Expression()). The function PMPLSFEM utilizes the Python package numpy, which is imported as np.

Step 1 (Define function space and bilinear form). Listing A.1 visualizes the definition of the finite element space V $= S^k(\mathcal{T}) \times RT_{k-1}(\mathcal{T})$, the inner product a $= a(\bullet, \bullet)$, and the functional L $\in X^*$ with L$(x) = -(f, \operatorname{div}\tau)_{L^2(\Omega)}$ for all $x = (v, \tau) \in X = H_0^1(\Omega) \times H(\operatorname{div}, \Omega)$.

```
RTk = FiniteElement('RT',mesh.ufl_cell(),polydegree)
Sk = FiniteElement('P',mesh.ufl_cell(),polydegree)
SkRTk = Sk*RTk
V = FunctionSpace(mesh,SkRTk)
u,sig = TrialFunctions(V)
v,tau = TestFunctions(V)
a = (inner(grad(u)-sig,grad(v)-tau) + div(sig)*div(tau))*dx
L = -f*div(tau)*dx
```

Listing A.1: Definition of V, a, and L in PMPLSFEM

Step 2–4 (Set boundary conditions, assemble system matrices, and solve linear system of equations). The FEniCS functions solve assembles the system matrices, includes boundary conditions (see [LMW12, Chap. 6] for more information on the assembly of system matrices and the inclusion of boundary conditions in FEniCS), and solves a linear system of equations to compute the solution $\mathfrak{u}_h = (u_h, \sigma_h) = (\mathrm{u\_h}, \mathrm{sigma\_h}) \in X_h = S_0^k(\mathcal{T}) \times RT_{k-1}(\mathcal{T})$ to the variational problem

$$a(\mathfrak{u}_h, x_h) = \mathrm{L}(x_h) \qquad \text{for all } x_h \in X_h. \tag{A.1}$$

Listing A.2 visualizes the application of the function solve with homogeneous Dirichlet boundary condition and "MUltifrontal Massively Parallel sparse direct Solver" (MUMPS).

```
uh_sh = Function(V)
bc = DirichletBC(V.sub(0), Constant(0.0), 'on_boundary')
solve(a == L, uh_sh, bc, solver_parameters={'linear_solver': 'mumps'})
SkSpace = FunctionSpace(mesh, Sk)
RTkSpace = FunctionSpace(mesh, RTk)
u_h = Function(SkSpace)
sigma_h = Function(RTkSpace)
assign(u_h, uh_sh.sub(0))
assign(sigma_h, uh_sh.sub(1))
```

Listing A.2: Computation of the solution to (A.1) in PMPLSFEM

Step 5 (Compute residual). Listing A.3 displays the computation of the piecewise constant function cell_residual $\in \mathbb{P}_0(\mathcal{T})$ with $\mathrm{cell\_residual}|_T = \eta^2(T) = \|\nabla u_h - \sigma_h\|^2_{L^2(T)} + \|f + \operatorname{div} \sigma_h\|^2_{L^2(T)}$ for all $T \in \mathcal{T}$.

```
DG0 = FunctionSpace(mesh, "DG", 0)
LocRes = TestFunction(DG0)
residual = (LocRes*(inner(grad(u_h)-sigma_h, grad(u_h)-sigma_h))+
            LocRes*(f+div(sigma_h))**2)*dx
cell_residual = Function(DG0)
assemble(residual, tensor=cell_residual.vector())
```

Listing A.3: Computation of the residual $\eta^2(T)$ for all $T \in \mathcal{T}$ in PMPLSFEM

Remark A.1.1 (LSFEM for Stokes)**.** *The FEniCS implementations of all LSFEM solvers are very similar to the implementation of PMPLSFEM, except the implementation of the Stokes solver StokesPseudostressLSFEM in SolverLSFEM.py. This solver adds a Lagrange multiplier $\lambda \in \mathbb{R}$, which ensures that the trace of the solution $\sigma_h \in RT_{k-1}(\mathcal{T}; \mathbb{R}^{2\times 2})$ with $k \in \mathbb{N}$ equals $0 = \operatorname{tr}(\sigma_h)$. More precisely, given a weight $\gamma > 0$, a polynomial degree $k =$ polydegree $\in \mathbb{N}$, and the identity matrix $I_{d\times d} \in \mathbb{R}^{d\times d}$, the function StokesPseudostressLSFEM defines the discrete space $V = S_0^k(\mathcal{T}; \mathbb{R}^d) \times RT_{k-1}(\mathcal{T}; \mathbb{R}^{d\times d}) \times \mathbb{R}$, the right-hand side $L \in V^*$, and the bilinear form $a : V \times V \to \mathbb{R}$ with, for all $(u_h, \sigma_h, \lambda), (v_h, \tau_h, \xi) \in V$,*

$$\begin{aligned} a(u_h, \sigma_h, \lambda; v_h, \tau_h, \xi) &= (\nabla u_h - \operatorname{dev} \sigma_h, \nabla v_h - \operatorname{dev} \tau_h)_{L^2(\Omega)} + \gamma\, (\operatorname{div} \sigma_h, \operatorname{div} \tau_h)_{L^2(\Omega)} \\ &\quad + \lambda\, (I_{d\times d}, \tau_h)_{L^2(\Omega)} + \xi\, (\sigma_h, I_{d\times d})_{L^2(\Omega)}, \\ L(v_h, \tau_h, \xi) &= -\gamma\, (f, \operatorname{div} \tau_h)_{L^2(\Omega)}. \end{aligned}$$

Then the function computes the solution $(u_h, \sigma_h, \lambda) \in V$ to

$$a(u_h, \sigma_h, \lambda; v_h, \tau_h, \xi) = L(v_h, \tau_h, \xi) \qquad \textit{for all } (v_h, \tau_h, \xi) \in V.$$

This variational problem is equivalent to the minimization of the least-squares functional (3.57) over the discrete space $X_h := S^k_0(\mathcal{T};\mathbb{R}^d) \times \{\tau_h \in RT_{k-1}(\mathcal{T};\mathbb{R}^{d\times d}) \mid \int \operatorname{tr}(\tau_h)\,\mathrm{d}x = 0\}$ *in the sense that* $(u_h,\sigma_h) = \arg\min_{x_h\in X_h} LS(f;x_h)$.

A.1.2 Computation of eigenvalues

The computation of the improved reliability constant $C(X_h)$ (see Algorithm 1 on page 23) and the upper bound $C_{\mathrm{LBB},h}$ for the LBB constant C_{LBB} (see Theorem 4.1.1) requires upper bounds $\mu^{\mathrm{up}}_{h,1},\dots,\mu^{\mathrm{up}}_{h,n}$ for the discrete eigenvalues $\mu_{h,1}\le\cdots\le\mu_{h,n}$ with $n\le\dim X_h$ from the eigenvalue problem (3.20), that is $\mu_{h,j}\le\mu^{\mathrm{up}}_{h,j}$ for all $j=1,\dots,n$. The computation of these upper bounds splits into the steps

1. assemble system matrices $A\in\mathbb{R}^{N\times N}$ and $B\in\mathbb{R}^{N\times N}$ and apply boundary conditions,
2. approximate eigenvectors $x_1,\dots,x_n\in\mathbb{R}^N\setminus\{0\}$ with $Ax_j=\mu_{h,j}Bx_j$ for $j=1,\dots,n$,
3. define $C=(C_{jk})^{k=1,\dots,n}_{j=1,\dots,N}\in\mathbb{R}^{N\times n}$ with rows $(C_{jk})_{j=1,\dots,N}=x_k$ for all $k=1,\dots,n$ and compute $A^{\mathrm{red}}=C^\top AC$ and $B^{\mathrm{red}}=C^\top BC$,
4. solve the eigenvalue problem $A^{\mathrm{red}}x=\mu_h B^{\mathrm{red}}x$.

The remainder of this section exemplifies these steps with the function LSevpPoisson in EVPsolverLSFEM.py. Input parameters are the triangulation $\mathcal{T}$, called mesh, the polynomial degree $k\in\mathbb{N}$ of the ansatz space $X_h=S^k_0(\mathcal{T})\times RT_{k-1}(\mathcal{T})$ from (3.40), called polydegree, and the number $n\le\dim X_h$, called nrEigs.

Step 1 (Assemble system matrices and include boundary conditions) The algorithm starts with the definition of the N-dimensional finite element space $\mathrm{V}=S^k(\mathcal{T})\times RT_{k-1}(\mathcal{T})$ with basis $(\phi_{h,1},\dots,\phi_{h,N})$ and the inner products $\mathrm{a}=a(\bullet,\bullet)$ and $\mathrm{b}=(\bullet,\bullet)_X$ from Section 3.2.1. Listing A.4 displays the code.

```
RTk = FiniteElement('RT',mesh.ufl_cell(),polydegree)
Sk = FiniteElement('P',mesh.ufl_cell(),polydegree)
SkRTk = Sk*RTk
V = FunctionSpace(mesh,SkRTk)
u,sig = TrialFunctions(V)
v,tau = TestFunctions(V)
a = (inner(grad(u)-sig,grad(v)-tau) + div(sig)*div(tau))*dx
b = (inner(grad(u),grad(v)) + inner(sig,tau) + div(sig)*div(tau))*dx
```

Listing A.4: Definition of V, a, and b in LSevpPoisson

Let $\mathcal{I}\subset\{1,\dots,N\}$ denote set of all indices $j=1,\dots,N$ which correspond to a basis function $\phi_{h,j}$ with degree of freedom on the Dirichlet boundary (see [LMW12, Chap. 3] for more details on the basis functions). The algorithm assembles the matrices $A=(A_{jk})_{j,k=1,\dots,N}$ and $B=(B_{jk})_{j,k=1,\dots,N}$ in $\mathbb{R}^{N\times N}$ with

$$A_{jk}=a(\phi_{h,j},\phi_{h,k}) \quad\text{and}\quad B_{jk}=(\phi_{h,j},\phi_{h,k})_X \qquad \text{for all } j,k\in\{1,\dots,N\}\setminus\mathcal{I}. \tag{A.2}$$

For all $j\in\mathcal{I}$ the j-th row and column in A equal the canonical basis vector $e_j\in\mathbb{R}^N$ and all entries in the j-th row and column in B equal zero. Listing A.5 displays the code for the computation of $A=\mathrm{A}$ and $B=\mathrm{B}$.

```
dummy = inner(Constant(1),v)*dx
bcs = DirichletBC(V.sub(0),Constant(0.0),"on_boundary")
A = PETScMatrix()
assemble_system(a,dummy,bcs,A_tensor=A)
B = PETScMatrix()
assemble_system(b,dummy,bcs,A_tensor=B)
bcs.zero(B)
```

Listing A.5: Computing A and B in LSevpPoisson

Step 2 (Compute eigenvectors). The algorithm computes the eigenvalues $\mu_{h,1} \leq \cdots \leq \mu_{h,n}$ from (3.20) by solving the equivalent eigenvalue problem: Given the matrices $A, B \in \mathbb{R}^{N\times N}$ from (A.2), seek $\mu_{h,j} \in \mathbb{R}$ and $x_j \in \mathbb{R}^N \setminus \{0\}$ with

$$Ax_j = \mu_{h,j} B x_j \qquad \text{with } x_j \in \mathbb{R}^N \text{ and } j = 1, \ldots, n. \tag{A.3}$$

The matrices A and B are symmetric and positive (semi-) definite. However, numerical difficulties result in non-symmetric matrices $\mathrm{A} \approx A$ and $\mathrm{B} \approx B$, that is $\mathrm{A} - \mathrm{A}^\top \neq 0 \neq \mathrm{B} - \mathrm{B}^\top$. Since the Krylov-Schur method of the SLEPc eigenvalue solver in FEniCS performs better for symmetric matrices [HRTV07, Sec. 2.3], the algorithm seeks the solution to (A.3) with $A := 1/2\,(\mathrm{A} - \mathrm{A}^\top)$ and $B := 1/2\,(\mathrm{B} - \mathrm{B}^\top)$. Listing A.6 displays the computation.

```
A_mat = as_backend_type(A).mat()
B_mat = as_backend_type(B).mat()
trans_mat = PETSc.Mat()
A_mat.transpose(trans_mat)
A_mat = 0.5*(A_mat+trans_mat)
A = PETScMatrix(A_mat)
B_mat.transpose(trans_mat)
B_mat = 0.5*(B_mat+trans_mat)
B = PETScMatrix(B_mat)
```

Listing A.6: Symmetrizing $A := 1/2\,(\mathrm{A} - \mathrm{A}^\top)$ and $B := 1/2\,(\mathrm{B} - \mathrm{B}^\top)$ in LSevpPoisson

Listing A.7 displays the code for solving (A.3) with the Krylov-Schur method of the SLEPc eigenvalue solver for symmetric matrices.

```
eigensolver = SLEPcEigenSolver(A,B)
eigensolver.parameters["solver"] = "krylov-schur"
eigensolver.parameters["problem_type"] = "gen_hermitian"
eigensolver.parameters["spectrum"] = "target magnitude"
eigensolver.parameters["spectral_transform"] = "shift-and-invert"
eigensolver.parameters["tolerance"] = 1e-25
eigensolver.parameters["spectral_shift"] = 0.0
eigensolver.solve(nrEigs)
```

Listing A.7: Computing $\tilde{\mu}_{h,1}, \ldots, \tilde{\mu}_{h,n}$ in LSevpPoisson

The iterative Krylov-Schur algorithm solves the eigenvalue problem (A.3) inexactly, that is, it approximates the exact eigenvalues $\mu_{h,1}, \ldots, \mu_{h,n}$ and eigenfunctions $\psi_{h,1}, \ldots, \psi_{h,n} \in S_0^k(\mathcal{T}) \times RT_{k-1}(\mathcal{T})$ by numbers $\tilde{\mu}_{h,1}, \ldots, \tilde{\mu}_{h,n} \in \mathbb{R}$ and functions $\tilde{\psi}_{h,1}, \ldots, \tilde{\psi}_{h,n} \in S^k(\mathcal{T}) \times RT_{k-1}(\mathcal{T})$. In general, there is a small error $|\tilde{\mu}_{h,j} - \mu_{h,j}| > 0$ for all $j = 1, \ldots, n$ and so the computation of $C(X_h)$ and C_{LBB} with $\tilde{\mu}_{h,1}, \ldots, \tilde{\mu}_{h,n}$ can result in lower bounds $\tilde{\mu}_{h,k} < \mu_{h,k}$ for some $k = 1, \ldots, n$. The computation of the reliability constant $C(X_h)$ requires upper eigenvalue bounds. These upper bounds result from Step 3–4.

Step 3 (Compute A^{red} and B^{red}). In general, the inexactly computed eigenfunctions from Step 2 do not satisfy the homogeneous Dirichlet boundary condition and so are not in the space X_h. The algorithm transforms the functions $\tilde{\psi}_{h,1}, \dots, \tilde{\psi}_{h,n} \notin X_h$ into functions $\hat{\psi}_{h,1}, \dots, \hat{\psi}_{h,n} \in X_h$ via the following routine. Let $\mathcal{I}$ be the index set from Step 1 and set

$$y_{jk} := \begin{cases} 0, & \text{if } j \in \mathcal{I}, \\ x_{jk}, & \text{if } j \in \{1, \dots, N\} \setminus \mathcal{I}. \end{cases} \qquad \text{with } \tilde{\psi}_{h,j} = \sum_{j=1}^{N} x_{jk} \phi_{h,k} \quad \text{for all } j = 1, \dots, n.$$

Then the function $\hat{\psi}_{h,j} := \sum_{j=1}^{N} y_{jk} \phi_{h,k} \in X_h$ for all $j = 1, \dots, n$. Listing A.8 displays the computation of the coefficient matrix EigFunc_MATRIX $= C = (y_{jk})_{j=1,\dots,N}^{k=1,\dots,n} \in \mathbb{R}^{N \times n}$.

```
dimX_h = B.size(0)
dimX_hMod = eigensolver.get_number_converged()
G_array = np.zeros([dimX_h,dimX_hMod])
row_array = np.zeros([dimX_h,dimX_hMod])
column_array = np.zeros([dimX_h,dimX_hMod])
bc_dict = bcs.get_boundary_values()
bd_Dofs = np.array([np.intc(dof) for dof in bc_dict])
for n in xrange(0,eigensolver.get_number_converged()):
  r1,c1,rx1,cx1 = eigensolver.get_eigenpair(n)
  r = rx1.get_local()
  r[bd_Dofs] = 0.0
  G_array[:,n] = r
  row_array[:,n] = xrange(0,dimX_h)
  column_array[:,n] = n*np.ones(dimX_h)
Atemp = csr_matrix((G_array.flatten(), (row_array.flatten(),
                    column_array.flatten())),shape=(dimX_h,dimX_hMod))
EigFunc_MATRIX = PETSc.Mat().createAIJ(size=(dimX_h,dimX_hMod),
                                       csr=(Atemp.indptr,Atemp.indices, Atemp.data))
```

Listing A.8: Computation of $C =$ EigFunc_MATRIX in LSevpPoisson

Listing A.9 displays the computation of the $n \times n$ matrices $A^{\text{red}} =$ Ared_mat and $B^{\text{red}} =$ Bred_mat with

$$A^{\text{red}} = (A_{jk}^{\text{red}})_{j,k=1,\dots,n} = C^\top A C \qquad \text{and} \qquad B^{\text{red}} = (B_{jk}^{\text{red}})_{j,k=1,\dots,n} = C^\top B C. \tag{A.4}$$

```
A_mat = as_backend_type(A).mat()
B_mat = as_backend_type(B).mat()
Ared_mat = PETSc.Mat()
Bred_mat = PETSc.Mat()
EigFunc_MATRIX.transposeMatMult(A_mat*EigFunc_MATRIX, result=Ared_mat)
EigFunc_MATRIX.transposeMatMult(B_mat*EigFunc_MATRIX, result=Bred_mat)
Ared_mat = Ared_mat.getValues(xrange(0,dimX_hMod),xrange(0,dimX_hMod))
Bred_mat = Bred_mat.getValues(xrange(0,dimX_hMod),xrange(0,dimX_hMod))
```

Listing A.9: Computation of $A^{\text{red}} =$ Ared_mat and $B^{\text{red}} =$ Bred_mat in LSevpPoisson

The matrices satisfy $A_{j,k}^{\text{red}} = a(\hat{\psi}_{h,j}, \hat{\psi}_{h,k})$ and $B_{j,k}^{\text{red}} = (\hat{\psi}_{h,j}, \hat{\psi}_{h,k})_X$ for all $j, k = 1, \dots, n$.

Step 4 (Solve eigenvalue problem). Listing A.10 displays the computation of eigenvalues $\mu_{h,1}^{\text{up}} \le \dots \le \mu_{h,n}^{\text{up}}$ with

$$A^{\text{red}} x_j = \mu_{h,j}^{\text{up}} B^{\text{red}} x_j \qquad \text{with } x_j \in \mathbb{R}^n \setminus \{0\} \text{ for all } j = 1, \dots, n. \tag{A.5}$$

Since the matrices in this generalized eigenvalue problem are small, the error of the algorithm eigvalsh from the scipy.linalg package is negligible, that is, the computation in Listing A.10 results in exact eigenvalues $\mu_{h,1}^{\mathrm{up}}, \dots, \mu_{h,n}^{\mathrm{up}}$.

```
Ared_mat = Ared_mat+Ared_mat.T
Bred_mat = Bred_mat+Bred_mat.T
eigVals = eigvalsh(Ared_mat,Bred_mat)
```

Listing A.10: Computation of $\mu_{h,1}^{\mathrm{up}}, \dots, \mu_{h,n}^{\mathrm{up}}$ in LSevpPoisson

The eigenvalues $\mu_{h,1}^{\mathrm{up}}, \dots, \mu_{h,n}^{\mathrm{up}}$ solve the eigenvalue problem: Seek $\mu_{h,j}^{\mathrm{up}} \in \mathbb{R}$ and $0 \neq \psi_{h,j}^{\mathrm{up}} \in X_h^{\mathrm{red}} := \mathrm{span}\{\hat{\psi}_{h,1}, \dots, \hat{\psi}_{h,n}\}$ with

$$a(\psi_{h,j}^{\mathrm{up}}, x_h) = \mu_{h,j}^{\mathrm{up}} (\psi_{h,j}^{\mathrm{up}}, x_h)_X \qquad \text{for all } x_h \in X_h^{\mathrm{red}} \text{ and } j = 1, \dots, n.$$

Since $X_{h,\mathrm{red}} \subset X_h$, the Rayleigh-Ritz principle yields $\mu_{h,j} \leq \mu_{h,j}^{\mathrm{up}}$ for all $j = 1, \dots, n$. In other words, the computed eigenvalues $\mu_{h,1}^{\mathrm{up}}, \dots, \mu_{h,n}^{\mathrm{up}}$ are guaranteed upper bounds for the discrete eigenvalues $\mu_{h,1}, \dots, \mu_{h,n}$.

Remark A.1.2 (Exact arithmetic)**.** *If the algorithm computes (A.2)–(A.5) exactly, the upper discrete eigenvalue bounds $\mu_{h,j}^{\mathrm{up}} = \mu_{h,j}$ for all $j = 1, \dots, n$.*

Remark A.1.3 (Inexact computation with FEniCS)**.** *The computation of the symmetric matrices A and B in Listing A.5 is inexact. For example, the Frobenius norm $\|A - A^\top\|_F = 9 \times 10^{-10} \neq 0$ for $N = 16641$, uniformly refined triangulation $\mathcal{T}$ of the square domain $\Omega = (0,1)^2$, and polynomial degree $k = 1$. These inexact computations might cause the numerical difficulties in the experiments of Section 3.2.1–3.2.3.*

A.2 Implementation of DPG

This section explains the computation of the solution $\mathfrak{u}_h \in X_h$ to the practical DPG method (5.3) in the numerical experiments of this thesis. It exemplifies the code with the function PMPprimalDPG in SolverDPG.py, which solves the primal DPG method for the Poisson model problem from Section 5.2.1. All solvers base on the software package FEniCS 2017.2.0. Their structure reads:

1. compute system matrices,
2. compute optimal test functions,
3. include boundary conditions,
4. solve linear system of equations,
5. compute residual.

Input parameter for all solvers in SolverDPG.py are a regular triangulation $\mathcal{T}$, called mesh (an element in the class dolfin.cpp.mesh.Mesh()), the right-hand side f (an element in the class dolfin.functions.expression.Expression()), and the integers polydegree and delta. Additional input parameters are possible, for example the weight rho and the frequency omega in HelmholtzPrimalDPG. Besides the software package FEniCS, the solvers utilize the package numpy (imported as np), the python bindings petsc4py for PETSc, and the matrix libraries numpy.matlib and scipy.sparse.

Step 1 (Compute system matrices). The solution space in Section 5.2.1 reads $X = H^1_0(\Omega) \times H^{-1/2}(\partial\mathcal{T})$ and the broken test space reads $Y = H^1(\mathcal{T})$. The solver circumvents the discretization of $H^{-1/2}(\partial\mathcal{T}) = \gamma_\nu^{\mathcal{T}} H(\mathrm{div}, \Omega)$ by discretizing the space $H(\mathrm{div}, \Omega)$ with the Raviart-Thomas finite element space $RT_k(\mathcal{T})$, $k \in \mathbb{N}_0$, from (3.40). FEniCS allows to remove all basis function that correspond to interior degrees of freedom with the command ['facet']. This avoids discrete functions $\tau_h \in RT_k(\mathcal{T})$ with $\gamma_\nu^{\mathcal{T}} \tau_h = 0$. Listing A.11 displays the discretization in PMPprimalDPG. The discretization involves the Courant space $S^k(\mathcal{T})$, the Raviart-Thomas space $RT_{k-1}(\mathcal{T})$, and the space of piece-wise polynomials $Y_h = \mathbb{P}_{k+\delta}(\mathcal{T})$, where $k, \delta \in \mathbb{N}_0$ equal the input parameter polydegree and delta.

```
RT_elem = FiniteElement('RT',mesh.ufl_cell(),polydegree)['facet']
S_elem = FiniteElement('P',mesh.ufl_cell(),polydegree)
Y_elem = FiniteElement('DG',mesh.ufl_cell(),polydegree+delta)
X_h = FunctionSpace(mesh,MixedElement([S_elem,RT_elem]))
Y_h = FunctionSpace(mesh,Y_elem)
```

Listing A.11: Discrete spaces in PMPprimalDPG

Let the basis of the m-dimensional space X_h = $X_h \subset H^1(\Omega) \times H^{-1/2}(\partial\mathcal{T})$ and the n-dimensional space Y_h = $Y_h \subset Y$ read $(x_1, x_2, \dots, x_m)$ and $(y_1, y_2, \dots, y_n)$. The solver computes the (sparse) system matrices B, G and the vector F with

$$B = (B_{jk})_{j=1,\dots,m}^{k=1,\dots,n} \in \mathbb{R}^{m\times n} \quad \text{with } B_{jk} = b(x_j, y_k) \text{ for all } j = 1, \dots, m, k = 1, \dots, n, \tag{A.6a}$$

$$G = (G_{jk})_{j=1,\dots,n}^{k=1,\dots,n} \in \mathbb{R}^{n\times n} \quad \text{with } G_{jk} = (y_j, y_k)_Y \text{ for all } j, k = 1, \dots, n, \tag{A.6b}$$

$$F = (F_j)_{j=1,\dots,n} \in \mathbb{R}^n \quad \text{with } F_j = (f, y_j)_{L^2(\Omega)} \text{ for all } j = 1, \dots, n. \tag{A.6c}$$

Listing A.12 displays the computation of B, G, and F in PMPprimalDPG.

```
y1 = TrialFunction(Y_h)
y2 = TestFunction(Y_h)
u,p = TrialFunctions(X_h)
a = (inner(grad(y1),grad(y2)) + y1*y2)*dx
b = (inner(grad(u),grad(y2)) - inner(p,grad(y2)) - div(p)*y2)*dx
rhs = (f*y2)*dx
dummy = Constant(0)*y2*dx
G = PETScMatrix()
B = PETScMatrix()
F = PETScVector()
assemble_system(a,rhs,A_tensor=G,b_tensor=F)
assemble_system(b,dummy,A_tensor=B)
```

Listing A.12: Computation of (A.6) in PMPprimalDPG

Step 2 (Compute optimal test functions). The computation of optimal test functions starts with the computation of the inverse $G^{-1} \in \mathbb{R}^{n\times n}$ of the matrix $G \in \mathbb{R}^{n\times n}$ from (A.6). The computation utilizes the fact that G is block diagonal with nrElem = $|\mathcal{T}|$ blocks. More precisely, the solver inverts all blocks (with the function np.linalg.inv), stores them, and assembles Ginv_mat = G^{-1}. Listing A.13 visualizes the code for the computation of G^{-1}.

```
MatrixDim = G.size(0)
nrElem = mesh.num_cells()
BlockSize = Y.dofmap().max_element_dofs()
G_array = np.zeros([BlockSize,BlockSize,nrElem])
row_array = np.zeros([BlockSize,BlockSize,nrElem])
```

```
column_array = np.zeros([BlockSize,BlockSize,nrElem])
for elem in xrange(0,nrElem):
  indices = range(elem*BlockSize,elem*BlockSize+BlockSize)
  G_array[:,:,elem] = np.linalg.inv(G.mat().getValues(indices,indices))
  column_array[:,:,elem] = np.matlib.repmat(indices,BlockSize,1)
  row_array[:,:,elem] = np.transpose(column_array[:,:,elem])
A = csr_matrix((G_array.flatten(), (row_array.flatten(),
                column_array.flatten())),shape=(MatrixDim,MatrixDim))
Ginv_mat = PETSc.Mat().createAIJ(size=(MatrixDim,MatrixDim),
                                 csr=(A.indptr,A.indices,A.data))
```

Listing A.13: Computation of Ginv_mat $= G^{-1}$ in PMPprimalDPG

Given a basis function $x_j \in X_h$ with $j \in \{1, \dots, m\}$, the optimal test function $T_h x_j \in Y_h$ satisfies $(T_h x_j, y_k)_Y = b(x_j, y_k)$ for all $k = 1, \dots, n$. Thus, the coefficients $\xi_1, \dots, \xi_n \in \mathbb{R}$ of the optimal test function $T_h x_j = \sum_{k=1}^{n} \xi_k y_k$ equal the j-th column of the matrix $G^{-1}B^\top$ with the transpose $B^\top$ of B from (A.6a) (see [RDM15, pp. 461–462] for more details). This and $G^{-1} = (G^{-1})^\top$ show that the coefficients $\zeta = (\zeta_j)_{j=1}^m \in \mathbb{R}^m$ of the solution $\mathfrak{u}_h = \sum_{j=1}^{m} \zeta_j x_j \in X_h$ to the practical DPG method $b(\mathfrak{u}_h, T_h x_j) = (f, T_h x_j)_{L^2(\Omega)}$ for all $j = 1, \dots, m$ solve the linear system of equations

$$BG^{-1}B^\top \zeta = BG^{-1}F.$$

Listing A.14 visualizes the code for the computation of the matrix K $= BG^{-1}B^\top$ and the vector RHS_vec $= BG^{-1}F$ in PMPprimalDPG.

```
B_mat = as_backend_type(B).mat()
F_vec = as_backend_type(F).vec()
K_mat = PETSc.Mat()
B_mat.transposeMatMult(Ginv_mat*B_mat, result=K_mat)
K = PETScMatrix(K_mat)
RHS_vec = as_backend_type(Function(X).vector()).vec()
temp_vec = F_vec.copy()
Ginv_mat.multTranspose(F_vec,temp_vec)
B_mat.multTranspose(temp_vec,RHS_vec)
RHS_vec = PETScVector(RHS_vec)
```

Listing A.14: Computation of K $= BG^{-1}B^\top$ and RHS_vec $= BG^{-1}F$ in PMPprimalDPG

Step 3 (Include boundary conditions). The basis $x_1, \dots, x_m$ of the discrete space X_h $\not\subset X$ from Listing A.11 contains functions which do not vanish on the Dirichlet boundary, that is, the index set $\mathcal{I} := \{j \in \{1, \dots, n\} \mid x_j \notin X\}$ is not empty. To include the boundary condition, the algorithm PMPprimalDPG replaces the j-th column in K $= BG^{-1}B^\top \in \mathbb{R}^{m \times m}$ by the canonical basis vector $e_j \in \mathbb{R}^m$ and the j-th entry in RHS_vec by zero for all $j \in \mathcal{I}$.

```
bc = DirichletBC(X_h.sub(0),Constant(0),'on_boundary')
bc_dict = bc.get_boundary_values()
bd_Dofs = np.array([np.intc(dof) for dof in bc_dict])
bd_Values = np.array([np.float64(bc_dict[dof]) for dof in bc_dict])
RHS_vec.set_local(bd_Values,bd_Dofs)
K.ident(bd_Dofs)
```

Listing A.15: Include Dirichlet boundary conditions in PMPprimalDPG

Step 4 (Solve linear system of equations). The algorithm solves the linear system of equations $K\zeta = RHS\_vec$ and transforms the coefficients vector $xVec = \zeta = (\zeta_j)_{j=1}^m \in \mathbb{R}^m$ into the function $(u_h, s_h) = \sum_{j=1}^m \zeta_j x_j \in X_h$. The solution to the DPG method (5.3) reads $\mathfrak{u}_h = (u_h, \gamma_\nu^{\mathcal{T}} s_h)$. Listing A.16 displays the FEniCS code.

```
x = Function(X_h)
xVec = x.vector()
solve(K, xVec, RHS_vec)
S = FunctionSpace(mesh, S_elem)
RT = FunctionSpace(mesh, RT_elem)
u_h = Function(S)
s_h = Function(RT)
assign(u_h, x.sub(0))
assign(s_h, x.sub(1))
```

Listing A.16: Computation of $(u_h, s_h) \in S_0^k(\mathcal{T}) \times RT_{k-1}(\mathcal{T})$ in PMPprimalDPG

Step 5 (Compute residual). Listing A.17 shows the FEniCS code which computes the coefficients $\xi = (\xi_j)_{j=1}^n = G^{-1}(B\zeta - F) \in \mathbb{R}^n$ of the Riesz representation $\eta_h = \sum_{j=1}^n \xi_j y_j \in Y_h$ of the functional $b(\mathfrak{u}_h, \bullet) - (f, \bullet)_{L^2(\Omega)} = (\eta_h, \bullet)_Y$ in Y_h.

```
eta_vec = as_backend_type(Function(Y_h).vector()).vec()
temp = eta_vec.copy()
xPETSc_vec = as_backend_type(xVec).vec()
B_mat.mult(xPETSc_vec, temp)
temp.axpy(-1, F_vec)
Ginv_mat.mult(temp, eta_vec)
```

Listing A.17: Computation of eta_vec $= \xi \in \mathbb{R}^n$ in PMPprimalDPG

The computable residual reads $\|\eta_h\|_Y^2 = \xi^\top G \xi$. The solver stores its local contributions $\eta^2(T) := \|\eta_h\|_{L^2(T)}^2 + \|\nabla \eta_h\|_{L^2(T)}^2$ for all $T \in \mathcal{T}$ in the $\mathbb{P}_0(\mathcal{T})$ function cell_residual. More precisely, the computation in Listing A.18 results in the function cell_residual $\in \mathbb{P}_0(\mathcal{T})$ with cell_residual$|_T = \eta^2(T)$ for all $T \in \mathcal{T}$.

```
DG0 = FunctionSpace(mesh, "DG", 0)
LocRes = TestFunction(DG0)
eta_vec = PETScVector(eta_vec)
eta = Function(Y_h, eta_vec)
residual = (LocRes*(inner(grad(eta), grad(eta)) + eta*eta))*dx
cell_residual = Function(DG0)
assemble(residual, tensor=cell_residual.vector())
```

Listing A.18: Computation of cell_residual in PMPprimalDPG

A.3 Further routines

A.3.1 Adaptive mesh refinement

Some experiments in this thesis utilize adaptive mesh refinements. More precisely, the experiments apply the loop

SOLVE, ESTIMATE, MARK, REFINE.

Algorithm 3: Adaptive mesh refinement (with Dörfler marking)

Input: bulk parameter $\Theta \in (0,1]$, triangulation $\mathcal{T}$, refinement indicator $(\eta^2(T))_{T\in\mathcal{T}} \in \mathbb{R}^{|\mathcal{T}|}$

1 MARK: choose a minimal subset $\mathcal{M} \subseteq \mathcal{T}$ with $\Theta \sum_{T\in\mathcal{T}} \eta^2(T) \leq \sum_{T\in\mathcal{M}} \eta^2(T)$;

2 REFINE: apply an adaptive mesh refinement, which refines all elements in $\mathcal{M}$ and results in a regular triangulation $\mathcal{T}_{\text{ref}}$;

Output: the refined triangulation $\mathcal{T}_{\text{ref}}$

Appendix A.1 and Appendix A.2 discuss the steps SOLVE and ESTIMATE. Algorithm 3 executes the steps MARK and REFINE. Listing A.19 visualizes the FEniCS function adaptiveRefinement from Mesh.py, which realizes Algorithm 3 in FEniCS. Input parameters are the triangulation $\mathcal{T}$ of the given domain $\Omega \subset \mathbb{R}^d$, called mesh, the bulk parameter Θ, called theta, and a function cell_residual $\in \mathbb{P}_0(\mathcal{T})$ with cell_residual$|_T = \eta^2(T)$ for all $T \in \mathcal{T}$. The numpy array ErrorIndicator equals $(\eta^2(T))_{T\in\mathcal{T}}$. The function cell_markers : $\mathcal{T} \to \{0,1\}$ corresponds to the set $\mathcal{M} \subset \mathcal{T}$ from Step 1 in Algorithm 3 in the sense cell_markers$(T) = 1$ for all $T \in \mathcal{M}$ and cell_markers$(T) = 0$ for all $T \in \mathcal{T} \setminus \mathcal{M}$. The new triangulation $\mathcal{T}_{\text{ref}}$ results from the FEniCS function refine(mesh,cell_markers). According to the DOLFIN User Manual (Feb. 24, 2006), the FEniCS routine refine applies the adaptive mesh refinement algorithm from [Bey95] with marked elements $\mathcal{M}$.

```
def adaptiveRefinement(mesh,theta,cell_residual):
  cell_markers = MeshFunction("bool",mesh,mesh.topology().dim())
  ErrorIndicator = np.array(cell_residual.vector())
  ErrorIndicatorTotal = np.sum(ErrorIndicator)
  idxSorted = np.argsort(ErrorIndicator)
  cell_markers.set_all(False)
  sumLocError = 0
  tempidx = mesh.num_cells()
  tempStopCriterion = theta*ErrorIndicatorTotal
  while sumLocError < tempStopCriterion:
    tempidx -= 1
    cell = Cell(mesh,idxSorted[tempidx])
    cell_markers[cell] = True
    sumLocError += ErrorIndicator[idxSorted[tempidx]]
  return(refine(mesh,cell_markers))
```

Listing A.19: Computation of $\mathcal{T}_{\text{ref}}$ with adaptiveRefinement from Mesh.py

Remark A.3.1 (Adaptive mesh refinement algorithms)**.** *An alternative to the adaptive mesh refinement algorithm from [Bey95] is the newest vertex bisection algorithm from [Mau95, Tra97]. The newest vertex bisection has been studied intensely (see for example [BDD04, GHS16, GSS14, Mit17, Ste08]). It allows for the proof of the discrete reliability in [CGS13a, Thm. 3.1]) and is, according to [CFPP14, p. 1202], the only mesh-refinement strategy known to fulfil the estimates in [CFPP14, Eq. 2.9–2.10]. These estimates are key ingredients in the proof of optimal convergence rates, see for example [CFPP14, Prop. 4.6 and 4.15] or [Ste07, Thm. 5.3].*

A.3.2 Lower eigenvalue bounds with Crouzeix-Raviart FEM

Guaranteed lower eigenvalue bounds for the Dirichlet eigenvalues of the Laplace operator from [CG14b] for the domain $\Omega \subset \mathbb{R}^2$ with regular triangulation $\mathcal{T}$ result in the lower eigenvalue bounds in Experiment 4 from Section 3.2.1. The computation in [CG14b] utilizes the Crouzeix-Raviart finite element space $CR_0^1(\mathcal{T}) := \{v_{CR} \in \mathbb{P}_1(\mathcal{T}) : v_{CR}$ is continuous at all midpoints of interior edges of $\mathcal{T}$ and $v_{CR} = 0$ at all midpoint of outer edges of $\mathcal{T}\}$ and the piecewise gradient $\nabla_{NC} : CR_0^1(\mathcal{T}) \to \mathbb{P}_0(\mathcal{T}; \mathbb{R}^2)$ with $(\nabla_{NC} v_{CR})|_T = \nabla v_{CR}|_T$ for all $v_{CR} \in CR_0^1(\mathcal{T})$ and $T \in \mathcal{T}$. This non-conforming space allows for the computation of eigenpairs $(\lambda_{CR,j}, \phi_{CR,j}) \in \mathbb{R} \times CR_0^1(\mathcal{T})$ for $j = 1, \dots, \dim CR_0^1(\mathcal{T})$ with $\lambda_{CR,1} \leq \lambda_{CR,2} \leq \dots$ in the sense that

$$(\nabla_{NC}\phi_{CR,j}, \nabla_{NC} v_{CR})_{L^2(\Omega)} = \lambda_{CR,j} \, (\phi_{CR,j}, v_{CR})_{L^2(\Omega)} \qquad \text{for all } v_{CR} \in CR_0^1(\mathcal{T}).$$

Let $h_{\max} > 0$ be the maximal mesh size of $\mathcal{T}$ and define the in [CG14a] improved value $\kappa = (1/48 + j_{1,1}^{-2})^{1/2} = 0.29823494289$ with the first root of the Bessel function $j_{1,1}$. Then [CG14b, Thm. 5.1] shows that the J-th exact Dirichlet eigenvalue λ_J of the Laplace operator satisfies

$$\frac{\lambda_{CR,J}}{1 + \kappa^2 h_{\max}^2 \lambda_{CR,J}} \leq \lambda_J. \tag{A.7}$$

This estimate enables the computation of lower eigenvalue bounds. Thereby, the following argumentation from the supplementary material of [CS18] omits the separation condition $h_{\max} \leq ((1 + 1/J)^{1/2}) - 1)/(\kappa \lambda_J^{1/2})$ from [CG14b, Thm. 5.1]. Let $\phi_1, \dots, \phi_J$ be eigenfunctions that correspond to the first J exact eigenvalues $\lambda_1 \leq \dots \leq \lambda_J$ and define the non-conforming interpolation operator $\mathcal{I}_{NC} : H_0^1(\Omega) \to CR_0^1(\mathcal{T})$ (see [CG14b, p. 2609] for the definition). If the functions $\mathcal{I}_{NC}\phi_1, \dots, \mathcal{I}_{NC}\phi_J \in CR_0^1(\mathcal{T})$ are linear independent, the argumentation from the *Proof of the lower bound in Theorem 5.1* in [CG14b] applies. If $\dim \operatorname{span}\{\mathcal{I}_{NC}\phi_1, \dots, \mathcal{I}_{NC}\phi_J\} < J$, then there exists some $v \in \operatorname{span}\{\phi_1, \dots, \phi_J\}$ with $\|v\|_{L^2(\Omega)} = 1$ and $\mathcal{I}_{NC} v = 0$. The interpolation error estimate $\|w - \mathcal{I}_{NC} w\|_{L^2(\Omega)} \leq \kappa h_{\max} \|\nabla_{NC}(w - \mathcal{I}_{NC} w)\|_{L^2(\Omega)}$ for all $w \in H_0^1(\Omega)$ [CG14a, Thm. 4] leads to

$$\begin{aligned} 1 = \|v - \mathcal{I}_{NC} v\|_{L^2(\Omega)}^2 &\leq \kappa^2 h_{\max}^2 \|\nabla_{NC}(v - \mathcal{I}_{NC} v)\|_{L^2(\Omega)}^2 \\ &= \kappa^2 h_{\max}^2 \|\nabla v\|_{L^2(\Omega)}^2 \leq \kappa^2 h_{\max}^2 \lambda_J. \end{aligned}$$

Since $\lambda_{CR,J}/(1 + \kappa^2 h_{\max}^2 \lambda_{CR,J}) \leq 1/(\kappa^2 h_{\max}^2)$, this results in (A.7).

A.4 Data medium containing the software

The FEniCS routines of this thesis are published online. These routines are provided under the terms of the GNU General Public License, version 3. Figure A.1 displays the routines. The content is briefly summarized as follows: The folder *Experiments* contains executable files which run the experiments of this thesis and the folder *Routines* provides all required functions. The functions and routines require FEniCS 2017.2.0 and Python 2.7.15, but easily extend to newer versions of FEniCS and Python. In addition, the functions and routines require the mshr package for FEniCS and the NumPy, Matplotlib, os.path, sys, and scipy libraries for Python as well as the python bindings petsc4py [BAA+18, DPKC11] for PETSc.

- Experiments
  - LSFEM_Poisson ..... runs Experiment 1–4 in Section 3.2.1
    - Exp1_PMP_asymptExact_knownSol.py
    - Exp2_PMP_asymptExact_unknownSol.py
    - Exp3_PMP_ImprovedGUB_Square.py
    - Exp4_PMP_ImprovedGUB_LShape.py
  - LSFEM_HelmMax ..... runs Experiment 1–4 in Section 3.2.2
    - Exp1_Helm_ImprovedGUBs.py
    - Exp2_Helm_Efficiency.py
    - Exp3_Helm_EstimatorCompetition.py
    - Exp4_Max_EstimatorCompetition.py
  - LSFEM_Stokes ..... runs the experiment in Section 3.2.3
    - Exp1_Stokes_asymptExactness.py
  - LSFEM_Clbb ..... runs Experiment 1–3 in Chapter 4
    - Exp1_CLBB_isolatedEVs.py
    - Exp2_CLBB_isolatedEV.py
    - Exp3_CLBB_Square.py
  - DPG_Helmholtz ..... runs Experiment 1–4 in Section 5.2.1
    - Exp1_unknownSol.py
    - Exp2_TraceNorm.py
    - Exp2b_TraceNorm.py
    - Exp3_Helm_asymptExact_knownSol.py
    - Exp4_DPG4HelmunknowSolInstantStability.py
  - DPG_Elasticity..... runs the experiment in Section 5.2.2
    - Exp1_Elast_Locking.py
    - Exp1b_Elast_localWeights.py
- Routines
  - DPGtools.py ..... approximation of trace norms
  - Mesh.py ..... geometries and the adaptive mesh refinement algorithm
  - Plot.py ..... log-log plot
  - EVPsolver ... eigenvalue solvers and computation of $C(X_h)$ from Section 3.1.2
    - EVPsolverGalerkin.py
    - EVPsolverNonConforming.py
    - EVPsolverLSFEM.py
  - Estimator ..... error estimator/indicator
    - Estimator.py
    - EstimatorDPG.py
    - EstimatorLS.py
  - Solver..... Galerkin, DPG, and LSFEM solver
    - SolverGalerkinFEM.py
    - SolverDPG.py
    - SolverLSFEM.py

Figure A.1: Content of the software archive